# Supplementary Material for "A genetic atlas of human admixture history"

Garrett Hellenthal, George B.J. Busby, Gavin Band,
James F. Wilson, Cristian Capelli, Daniel Falush,
& Simon Myers

## Contents

# List of Tables



# List of Figures

# S1 Précis of admixture inference procedure

Our approach is based on characterising sampled admixed populations that contain a mixture of ancestries from sources related to different sampled donor groups, using properties of the joint probabilities of donor ancestries at positions separated by different genetic distances $g$. These probabilities are derived, for a variety of different admixture scenarios we study, in Note S2. Relying on this, we describe the basis of our method: chromosome painting, and properties of painted chromosomes, in Note S3. In Note S4, we describe the nuts and bolts of our inference procedure. To avoid confusion, we refer to "sources" as the true (unsampled) admixing source groups, and we refer to sampled groups in our dataset as "recipients" or "donors" (i.e. under our model).

This précis broadly summarizes the steps in our inference procedure, by which admixture is inferred in a recipient population using a collection of $K-1$ donor populations, indicating which sections of the supplement are relevant to understanding each step. Although there are a number of steps involved in the approach, several of these relate to the need to test for the diverse possibilities for different modes of admixture (many of which we frequently infer), including no admixture at all, simple admixture between two groups at a single time, admixture at multiple times, admixture involving three or more groups, and to consider potential, even more complex, models including more continuous admixture or admixture involving many groups.

## Painting

1. We infer a "copying vector" for each population in the dataset by performing chromosome painting, representing each recipient individual's genome as a mosaic of the genomes of the donor population individuals, as described in Note S3.1. This "copying vector" is of length $K$, with elements giving the proportion of genome-wide DNA that the recipient individuals copy from each of the $K$ groups. The expected properties of "copying vectors" of admixed groups are derived in Note S3, with properties that we assume for the painting in deriving these results listed in Note S3.1. The details of the painting protocol are described in Note S4.1.1. In this painting, we allow each of the $K$ populations to copy from every other population including itself.

2. We generate 10 "painting samples" for each chromosome of a recipient individual using the detailed protocol described in Note S4.1.2. (Ten is an arbitrary number, designed to capture variability in painting realisations, which seems to work well in practice.) Apart from actually sampling particular paintings along the genome, this step differs from the previous step in that to avoid masking admixture signals, here copying from members with the same population label is NOT allowed.

## Initial mixture modelling of the recipient population and "cleaned painting"

3. We generate an initial representation of the haplotype composition of the recipient population by modelling it as a mixture of the haplotypes from the other (donor) populations in the sample (Note S3.2). Specifically, we perform a non-negative-least-squares (nnls) regression, taking the copying vector of the recipient population (after subtracting out the contribution from individuals with the same population label) as the response and the copying vectors for each donor population (after subtracting out the contribution from the recipient population) as the predictors. The coefficients of this regression are restricted

to be $\geq 0$ and to sum to 1 across donors, and we describe how this is performed in detail in Note S4.2. This mixture model typically has many fewer than $K-1$ populations with non-zero contributions.

4. Using the mixture model and initial painting, we generate a "cleaned painting" (Note S3.3) for each of the 10 "painting samples" per each chromosome of a recipient individual, by reweighting segments coming from all $K-1$ donor groups, such that non-zero weights are only assigned for donor groups contributing to the mixture representation from step 3. For clarity, we note that the cleaned painting consists of a vector of $K-1$ weights at each position in the genome (typically with weights equal to zero for most of the $K-1$ populations), unlike the "painting sample" which consists of a single assignment of one of $K-1$ populations at each position.

## Generating coancestry curves

5. We use the "cleaned painting" to generate empirical coancestry curves using the protocol described in Note S4.3. Within and between every pairing of 10 "painting samples" generated for the two haploids of a recipient individual, we consider every pair of "chunks" (i.e. individual DNA segments in the mosaic representation) separated by genetic distance $g$. For any possible pair of donor groups contributing to the mixture representation, the left and right chunks each give a weight to these respective groups. We measure the average product of weights, at separation distance $g$, relative to the product of genome-wide average weights for the same pair, to form a coancestry curve as $g$ varies. These coancestry curves illustrate the decay in ancestry linkage disequilibrium versus genetic distance, so capture key details of the admixture history under different scenarios, and this is shown in Notes S3.4-S3.7. Summing over both haploids of the same individual naturally accounts for phasing "switch errors", a common source of error when inferring haplotypes. There is one such curve for each pair of donor populations in the mixture representation – all of these curves are provided for this analysis at http://admixturemap.paintmychromosomes.com/.

## Initial fit of admixture between two groups, at a single time

6. For admixture models with a single date involving only two groups, Note S3.5 shows how coancestry curves and copying vectors relate to admixture details. In particular, each coancestry curve is predicted to decay at the same rate $\lambda$, the number of generations since admixture. To estimate admixture time, we therefore find the maximum likelihood estimate (MLE) of rate parameter $\lambda$ of an exponential distribution fit to all coancestry curves simultaneously, using a specific procedure described in Note S4.4. This procedure returns an estimate of the admixture time $\lambda$ in generations, but also specific values for the intercept of each coancestry curve, forming an intercept "matrix" across all pairs of groups involved in the mixture. We use bootstrapping to generate approximate 95% CIs for $\lambda$.

7. To fit the source groups involved in admixture, we model the haplotypes of each true admixing source A and B as a mixture of those of sampled (donor) groups, and try to infer the mixture components and the admixture proportion $\alpha$. Notes S3.2 and S3.5 explain how the copying vectors and intercept matrix, respectively, relate to the mixture coefficients and $\alpha$ under this model. In particular, the intercept matrix has predicted rank 1 in this simple admixture scenario, and its first (and only) eigenvector relates

linearly to the source group mixture coefficients, as does the overall copying vector for the admixed group. We thus eigendecompose our estimated intercept matrix (in a manner similar to that employed in principal components analysis in genetics), and together with the estimated copying vector, we estimate both $\alpha$ and the mixture components for each true admixing source using a non-negative least squares approach so as to minimise the weighted summed squared differences between observed and predicted values. The details of this step are described in Note S4.5.

8. This fitting step re-estimates the mixing coefficients for the recipient population as a whole to be $\hat{\alpha}$ (i.e. the MLE of $\alpha$) times the inferred mixing coefficients of the first source plus $1-\hat{\alpha}$ times the inferred mixing coefficients of the second source, resulting in a new cleaned painting, and new coancestry curves, allowing another round of the fitting step itself. We therefore iterate the (re)estimation of the admixture date, and ancestral source population composition 5 times in total. This results in final estimates of $\alpha$, $\lambda$, and a representation of the haplotypes within each source group as found in a weighted mixture of sampled donor groups. This potentially characterizes admixture, if it is simple. However, first we test for evidence of admixture, and then for evidence of more complex admixture (using a slightly different characterisation of events in complex cases), as described in detail in Note S4.6.

## Testing for admixture and classifying the admixture event

9. In order to test for the presence of admixture, we first renormalize the coancestry curves to allow for variation in ancestry informativeness along the genome (see Note S4.7). Under our model, these normalized curves should behave exactly as the original curves, but in practice they are expected to be more robust to modelling departures, although noisier due to the normalisation, as shown in e.g. a setting of strong population bottlenecks considered in Note S5.5.3. We re-estimate the two source populations now using these normalized curves, and again use bootstrapping to generate approximate 95% CIs for the time since admixture $\lambda$, with 100 bootstrap samples. We note that inferred dates that are $= 1$ or $\geq 400$ correspond to no correlation of ancestry with distance (very old admixture means we expect ancestry chunks to be undetectably short so gives the same signal as no admixture). An empirical $p$-value testing the null of no admixture is thus given by the proportion of bootstrapped (and point estimate) dates that are $= 1$ or $\geq 400$, and we reject a null of no admixture only if $p < 0.01$. If the null is rejected, we also check for consistency of inferred admixture time with the previous analysis described in step 6, classifying admixture as "uncertain" (i.e. not characterizable) if there is no overlap in the 95% CIs.

10. If the analysis using normalized curves finds evidence of admixture, then admixture might be at one time or more than one time. If the latter, we show in Note S3.6 that we expect coancestry curves to be a mixture of exponential distributions with decay rates the admixture times, rather than showing a simple exponential decay. In practice we only examine models with up to two such decay rates, because we do not believe there is often power to determine additional admixture times (Notes S2.4 and S3.7), and interpret such fits as indicating admixture simply at "more than one time". Further the components of the intercept matrices corresponding to each decay rate allow partial characterisation of the admixing group(s). We thus test for evidence of multiple dates based on the maximum improvement in fit-quality (measured by coefficient-of-determination) for fitting the coancestry curves as a mixture of exponentials relative to a single exponential decay

rate, inferring multiple dates if the empirical $p$-value of this test is $<0.05$. In this case, we attempt to learn partial information about the admixing sources, based on the first eigenvectors of the respective intercept matrix components for each decay rate, but we are not able to infer these sources completely. This procedure is described in Note S4.8.

11. If the analysis finds evidence of admixture, but no evidence this is at more than one time, then admixture at the same time might involve $L$ source groups, where $L \geq 2$. In general, the intercept matrix estimated from the coancestry curve is predicted to have rank $L-1$ in this case (Note S3.6). We only attempt to analyse cases where $L \leq 3$, so classify admixture as "uncertain" if there is evidence the rank of our estimated intercept matrix is 3 or greater. Otherwise, we reject a null of simple 2-way admixture if there is empirical evidence $p < 0.05$ that the intercept matrix has rank 2, in which case we classify admixture as "multiway". In this case, we attempt to learn partial information about the admixing sources, based on both the first and second eigenvector of the intercept matrix, but we are not able to infer these sources completely. Details of this test are found in Note S4.9. Otherwise, admixture is classified as (being consistent with) a simple single event, and we use the existing inference (i.e. from steps 6-8) of groups involved.

# S2 Pairwise ancestry decay curves under single, multiple and continuous admixture models

In this note we consider the distribution of ancestry chunk sizes under several models of admixture. The results obtained underpin all our approaches, including the methods we develop for identifying groups with no admixture, a single admixture event involving two groups, more complex events with multiple groups, and more complex events with multiple admixture times, in real data.

We model recombination events in each generation as occurring as a Poisson process of rate 1 per Morgan in genetic distance along the genome. This ignores crossover interference, but interference is likely to be a minor factor after multiple generations following an admixture event.

## S2.1 Fundamental notation and assumptions

We describe here a series of assumptions about the nature of admixture events that are used throughout in our approach. Additional assumptions relating to the GLOBETROTTER approach itself are detailed in Note S3.1. We test the effect of these assumptions and robustness to them in Notes S5 and S6.4. In general we consider admixture between source groups labeled $S_1, S_2, ..., S_L$. We will use $S_1 = A, S_2 = B$, etc. to denote these (unobserved) admixing groups. In contrast, to avoid confusion we will refer to our $K$ sampled populations as labeled by $1, ..., K$ throughout this supplement. In an admixed population formed by mixing of these source groups $S_v$ some time(s) in the past, individuals have genomes consisting of segments of ancestry from the contributing groups. Suppose a source group labeled $S_v$ contributes a fraction $p_{S_v}$ of ancestry (i.e. of the genome), where $\sum_{v=1}^{L} p_{S_v} = 1$. Given these $p_{S_v}$, which specify the single-locus ancestry proportions, a natural question is to study the pairwise ancestry distribution of two positions a genetic distance $g$ apart. Specifically, define $p_{S_vS_w}(g)$ to be the probability that two such sites have ancestry $S_v, S_w$ respectively.

All of our work for this manuscript relies fundamentally on properties of $p_{S_vS_w}(g)$ as $v, w$ and $g$ vary. We use this group of functions to characterize properties of the admixture events, and their relative time(s) in the past. Intuitively, because ancestry chunks have non-zero size, nearby pairs of positions, corresponding to small $g$, are likely to have the same ancestry, while distant pairs will behave essentially independently. The scale of this ancestry correlation along the genome is captured by the variation of $p_{S_vS_w}(g)$ with $g$. The more ancient the admixture, the smaller the ancestry chunks, and so the more rapidly $p_{S_vS_w}(g)$ decays towards background as $g$ increases, allowing dating of admixture events.

We assume throughout that the admixed population has been random mating since admixture (if any) began. Further, we assume that since admixture the population size has been sufficiently large, relative to the time since admixture, that recent coalescence events – which could generate long-range linkage disequilibrium influencing ancestry segments – can be neglected, an assumption whose impact we later test in simulations (see Note S5). We also assume that ancestry proportions are on average uniform across the genome (i.e. no selection has occurred).

The theory in this section partially overlaps that of previous work (11; 41) but is included for completeness, and to define a consistent notation in this work.

## S2.2 Single admixture event

Consider the most simple setting, where a single admixture event occurs between two groups labeled $A$ and $B$ at a specific time $\lambda$ (assumed discrete) generations of recombination in the

past (much of our work corresponds to this case, but we also consider more complex settings; see below). Suppose the fraction of ancestors from population $A$ is $p_A = \alpha$, so $p_B = 1 - \alpha$. The assumptions of the previous section imply that distinct ancestry segments are independently drawn from population $A$ with probability $\alpha$.

Then the probability of no recombination between two points a distance $g$ apart since admixture is $\exp^{-g\lambda}$ and conditioning on whether such recombination occurred, we have (1):

$$p_{AA}(g) = \alpha(\exp^{-g\lambda} + [1 - \exp^{-g\lambda}]\alpha) = \alpha^2 + \alpha(1-\alpha)\exp^{-g\lambda}.$$

Now using identities of the form $\alpha = p_A = p_{AA}(g) + p_{AB}(g)$ and $p_B = p_{AB}(g) + p_{BB}(g)$ we immediately have:

$$\begin{aligned} p_{AB}(g) &= p_{BA}(g) = \alpha(1-\alpha) - \alpha(1-\alpha)\exp^{-g\lambda} \\ p_{BB}(g) &= (1-\alpha)^2 + \alpha(1-\alpha)\exp^{-g\lambda}. \end{aligned} \tag{S1}$$

So viewed as functions of genetic distance $g$, $p_{AA}(g)$ and $p_{BB}(g)$ decay exponentially towards their expectations under independence as $g$ increases, with a decay rate (in genetic distance) given by the time in generations since admixture. This property is the key to our dating approach, since it implies fitting of exponential curves can date admixture events. Note furthermore that $p_{AB}(g)$ *increases* with $g$ towards its independence expectation.

Note that knowledge of (any of) these curves completely characterize(s) the admixture event in this setting. We can use the simple form of the curves in the case of a single historical admixture event in order to test whether data are compatible with this model, or imply a more complex scenario.

## S2.3 Double admixture event

The behavior of admixture chunks under more complex settings is exemplified by the case of two admixture events.

Specifically, we consider a setting where there are two populations labelled $A$ and $B$ that admixed a total of $\lambda_1$ generations ago in a simple admixture event, to produce an admixed group, and then this group admixed a second time, with a third population ($C$), $\lambda_2$ generations ago. The initial event has admixture proportions $\alpha_1$ and $\alpha_2 = 1 - \alpha_1$, while the second event has proportion $\alpha_3$ for the third population. Hence today, the ancestry contributions of groups $A$, $B$ and $C$ respectively are $p_A = \alpha_1(1-\alpha_3)$, $p_B = \alpha_2(1-\alpha_3)$, and $p_C = \alpha_3$.

As before, assuming random mating of individuals between admixture events, we can extend the previous argument and find:

$$\begin{aligned} p_{AA}(g) &= \alpha_1^2(1-\alpha_3)^2 + \alpha_1^2\alpha_3(1-\alpha_3)\exp^{-g\lambda_2} + \alpha_1\alpha_2(1-\alpha_3)\exp^{-g\lambda_1} \\ p_{BB}(g) &= \alpha_2^2(1-\alpha_3)^2 + \alpha_2^2\alpha_3(1-\alpha_3)\exp^{-g\lambda_2} + \alpha_1\alpha_2(1-\alpha_3)\exp^{-g\lambda_1} \\ p_{AB}(g) &= p_{BA}(g) = \alpha_1\alpha_2(1-\alpha_3)^2 + \alpha_1\alpha_2\alpha_3(1-\alpha_3)\exp^{-g\lambda_2} - \alpha_1\alpha_2(1-\alpha_3)\exp^{-g\lambda_1} \\ p_{CA}(g) &= p_{CA}(g) = \alpha_3(1-\alpha_3)\alpha_1 - \alpha_3(1-\alpha_3)\alpha_1\exp^{-g\lambda_2} \\ p_{CB}(g) &= p_{BC}(g) = \alpha_3(1-\alpha_3)\alpha_2 - \alpha_3(1-\alpha_3)\alpha_2\exp^{-g\lambda_2} \\ p_{CC}(g) &= \alpha_3^2 + \alpha_3(1-\alpha_3)\exp^{-g\lambda_2}. \end{aligned} \tag{S2}$$

Notice that, for each ancestry combination, the decay (or increase) as genetic distance becomes large is the probability as calculated under independence, while the decay curve to this point is a mixture of exponentials with rate parameters $\lambda_1$ and $\lambda_2$, the times since admixture. Further, the decay curves involving the most-recently introduced group $C$ are simple exponential curves with rate parameter $\lambda_2$, the arrival time of this group. Thus, knowledge of these curves again in principle allows full characterization of the admixture event.

Two special cases of interest, and which are historically plausible in the case of humans, are those where two of the groups are identical (without loss of generality call these identical groups $B$ and $C$, so the same group admixes twice, for example due to repeated migrations), and where the two admixture times are identical (so three groups admix simultaneously). In the first case there are only two distinct ancestries $A$ and $B$, as in the single-admixture case considered above. Analogously, we have ancestry proportions: $p_A = \alpha = \alpha_1(1-\alpha_3)$, $p_B = 1-\alpha$. We simply sum the terms corresponding to groups $B$ and $C$ above and simplify to give:

$$\begin{aligned}
p_{AA}(g) &= \alpha^2 + \alpha(\alpha_1 - \alpha)\exp^{-g\lambda_2} + \alpha(1-\alpha_1)\exp^{-g\lambda_1} \\
p_{AB}(g) &= p_{BA}(g) = \alpha(1-\alpha) - \alpha(\alpha_1 - \alpha)\exp^{-g\lambda_2} - \alpha(1-\alpha_1)\exp^{-g\lambda_1} \\
p_{BB}(g) &= (1-\alpha)^2 + \alpha(\alpha_1 - \alpha)\exp^{-g\lambda_2} + \alpha(1-\alpha_1)\exp^{-g\lambda_1}\,.
\end{aligned}$$

If we define weights (corresponding to the increase of the fraction of ancestry from group $B$ at the successive admixture events, relative to the current day fraction)

$$w_1 = \frac{1-\alpha_1}{1-\alpha}, w_2 = \frac{\alpha_1 - \alpha}{1-\alpha},$$

then we have

$$\sum_{j=1}^{2} w_j = 1,$$

and we may write

$$\begin{aligned}
p_{AA}(g) &= \alpha^2 + \alpha(1-\alpha)\textstyle\sum_{j=1}^{2} w_j \exp^{-\lambda_j g} \\
p_{AB}(g) &= p_{BA}(g) = \alpha(1-\alpha) - \alpha(1-\alpha)\textstyle\sum_{j=1}^{2} w_j \exp^{-\lambda_j g} \\
p_{BB}(g) &= (1-\alpha)^2 + \alpha(1-\alpha)\textstyle\sum_{j=1}^{2} w_j \exp^{-\lambda_j g}\,.
\end{aligned} \tag{S3}$$

Thus, two distinct episodes of admixture involving the same group can be distinguished from a single admixture event by the fact that the joint ancestry curves do not have a simple exponential decay, but instead a weighted mixture of exponential decay rates.

In the second case, of three groups admixing simultaneously, it is natural to reparametrise $p_A = \beta_A = \alpha_1(1-\alpha_3)$, $p_B = \beta_B = \alpha_2(1-\alpha_3)$ and $p_C = \beta_C = \alpha_3$ and Equation S2 still holds with $\lambda_1 = \lambda_2 = \lambda$, giving after simplifying again:

$$\begin{aligned}
p_{AA}(g) &= \beta_A^2 + \beta_A(1-\beta_A)\exp^{-g\lambda} \\
p_{BB}(g) &= \beta_B^2 + \beta_B(1-\beta_B)\exp^{-g\lambda} \\
p_{AB}(g) &= p_{BA}(g) = \beta_A\beta_B - \beta_A\beta_B\exp^{-g\lambda} \\
p_{AC}(g) &= p_{CA}(g) = \beta_A\beta_C - \beta_A\beta_C\exp^{-g\lambda} \\
p_{BC}(g) &= p_{CB}(g) = \beta_B\beta_C - \beta_B\beta_C\exp^{-g\lambda} \\
p_{CC}(g) &= \beta_C^2 + \beta_C(1-\beta_C)\exp^{-g\lambda}\,.
\end{aligned} \tag{S4}$$

Note that here knowledge of any two of the curves $p_{S_vS_w}(g)$ would characterize the admixture event, and all curves decay (or increase) exponentially with a single rate parameter given by the admixture time, so fitting this single rate will give the time of the (multi-way) admixture event. It is straightforward to consider more complex scenarios of 3-way and more general admixture events, some of which may occur simultaneously; the formulas in this setting are not given, and are more complex. However, in general the $p_{S_vS_w}(g)$ follow the same form, with the joint ancestry probabilities always expressable as mixtures of exponentials, and the decay rates in the exponential mixtures corresponding to the admixture times involved. An important special case though, which we believe is likely to be common in real datasets, is continuous admixture, which we address in the next section.

In general, the results in this section imply identification of $p_{S_vS_w}(g)$ is sufficient to date admixture event(s), by fitting exponential decay rate(s), and to identify the admixture proportions involved, which we view as solving the problem of characterizing the events. If these joint ancestry curves fit a single exponential model, this implies admixture at a single point in the past; this can then be used as a diagnostic for more complex events. In practice, we will show that although we do not directly observe ancestry in real data, we can still use these ideas to perform inference of both the admixture time(s), and the underlying groups, by constructing quantities analogous to $p_{S_vS_w}(g)$ directly from the data.

An important caveat to this approach, which we explore in the next section for the continuous admixture case, is the complexity of dissecting mixtures of exponential curves, which can be a difficult problem. Specifically, mixtures of exponentials can appear to relatively closely fit a single exponential curve, as we see in the next section, and distinguishing, e.g. 3 from 4 embedded decay rates within a single curve is even more problematic. In practice, this is likely to mean our power to distinguish a single "pulse" of admixture from a range of admixture times is incomplete, particularly if the time range is relatively narrow.

## S2.4 Continuous admixture

The last setting we explicitly analyze is the perhaps realistic scenario where admixture occurs "continuously" over some time range. Specifically, we assume that an initial population labeled $A$ continually receives migrants from a second group $B$ (which does not alter) starting at time $\lambda_s$ and ending at time $\lambda_e$ generations in the past. For simplicity, we assume the admixture rate is constant, so that each generation a fraction $\mu$ of the population is derived of individuals from population $B$. It is immediate that in the present day, the fraction of genetic material from group $A$ is $p_A = \alpha = (1-\mu)^{\lambda_s - \lambda_e + 1}$. If the rate of migration varies through time, it is straightforward to adapt our results, but the broad conclusions remain unchanged.

The distribution of admixture chunk sizes here is specified by the probabilities $p_{S_vS_w}(g)$ for $S_v, S_w = A$ or $B$. We only need to derive $p_{AA}(g)$; other terms are straightforward from this. For two loci a distance $g$ apart, we condition on the time in the past, in generations, of the most recent recombination between them. This might be more recent than time $\lambda_e$, in some generation $j$ where $\lambda_e < j \leq \lambda_s$, or older than generation $\lambda_s$. In general, the probability the most recent recombination occurred $j$ generations in the past is given by $\Pr(j) = \exp^{-(j-1)g}(1 - \exp^{-g})$, independently of ancestry at the left locus. $p_{AA}(g)$ is the probability that the left locus has ancestry 1, and so does the right hand locus; thus we have:

$$\begin{aligned} p_{AA}(g) \;=\; & p_A \textstyle\sum_{j=1}^{\infty} \big[ \Pr(j) \\ & \Pr(\text{right locus has ancestry } A \mid \text{left locus anc. } A, \text{recombination in generation } j) \big]. \end{aligned}$$

In the generation where recombination occurs, the right hand locus is of ancestry $A$ with the probability an individual in that generation carries this ancestry. For example, this probability is 1 if $j > \lambda_s$ and $(1-\mu)^{\lambda_s - \lambda_e + 1}$ if $j \leq \lambda_e$. Summing over the possibilities gives:

$$\begin{aligned} p_{AA}(g) = (1-\mu)^{\lambda_e - \lambda_s + 1} \Big[ & (1 - \exp^{-g\lambda_e})(1-\mu)^{\lambda_s - \lambda_e + 1} \\ & + \textstyle\sum_{j=\lambda_e+1}^{\lambda_s} (1-\mu)^{\lambda_s - j + 1} \exp^{-g(j-1)}(1 - \exp^{-g}) + \exp^{-g\lambda_s} \Big], \end{aligned}$$

and simplifying

$$p_{AA}(g) = p_1^2 + \mu(1-\mu)^{\lambda_s - \lambda_e + 1} \sum_{j=\lambda_e}^{\lambda_s} (1-\mu)^{\lambda_s - j} \exp^{-gj}.$$

(Although this could be further simplified, we feel it is intuitive to represent the result as a mixture of exponentials.) We can now immediately derive the other $p_{S_v S_w}(g)$ as in the case of a single admixture event:

$$p_{AB}(g) = p_{BA}(g) = p_A p_B + \mu(1-\mu)^{\lambda_s - \lambda_e + 1} \sum_{j=\lambda_e}^{\lambda_s} (1-\mu)^{\lambda_s - j} \exp^{-gj}.$$
$$p_{BB}(g) = p_B^2 + \mu(1-\mu)^{\lambda_s - \lambda_e + 1} \sum_{j=\lambda_e}^{\lambda_s} (1-\mu)^{\lambda_s - j} \exp^{-gj}.$$

If we define weights

$$w_j = \frac{\mu(1-\mu)^{\lambda_s - j}}{1 - (1-\mu)^{\lambda_s - \lambda_e + 1}},$$

then recalling $\alpha = (1-\mu)^{\lambda_s - \lambda_e + 1} = p_A$, we have

$$\sum_{j=\lambda_e}^{\lambda_s} w_j = 1$$

and

$$p_{AA}(g) = \alpha^2 + \alpha(1-\alpha) \sum_{j=\lambda_e}^{\lambda_s} w_j \exp^{-gj}, \tag{S5}$$

which is the same form as for the two admixture event case (we can regard the latter as a special case of non-uniform continuous admixture). The other $p_{S_v S_w}(g)$ terms have analogous representations. Thus in general, as one might intuitively expect, the terms $p_{S_v S_w}(g)$ are mixtures over weighted individual exponential curves with decay rates in the range $[\lambda_e, \lambda_s]$.

We present example curves in this setting (Figure S1) for continuous admixture events between (i) 20 and 40 generations in the past and (ii) 20 and 100 generations in the past. Visually, in case (i) the curves show a decay rate intermediate between 20 and 40 generations, and in fact look quite well fit by a simple exponential decay curve, corresponding to a single admixture event of an age intermediate between these times, with 30 generations giving the best single date fit to Equation S1. In contrast, for (ii) the very broad range of admixture times means a single event provides a less good fit, with 50 generations giving the best single date fit. In both cases, the residuals from fitting a single admixture event show a specific pattern as $g$ varies (Figure S1), suggesting a means to identify continuous admixture after attempting fitting a single decay rate to the data. However, the absolute magnitude of these residuals is small, and we believe that, in practice, very large datasets will be required to distinguish continuous admixture from single admixture events, unless continuous admixture occurs over time ranges that are large relative to the average admixture age.

Even more challenging will be distinguishing, for example, continuous admixture and the occurrence of two distinct admixture pulses. One positive result is that fitting a single date gives answers that appear "sensible" in that the value obtained is close, in the examples, to the true mean admixture time (30 generations, and 60 generations, respectively). In addition, note that this problem only applies to the case where the migrating group remains genetically unchanged through time; waves of genetically distinct groups are likely to be more easily distinguished.

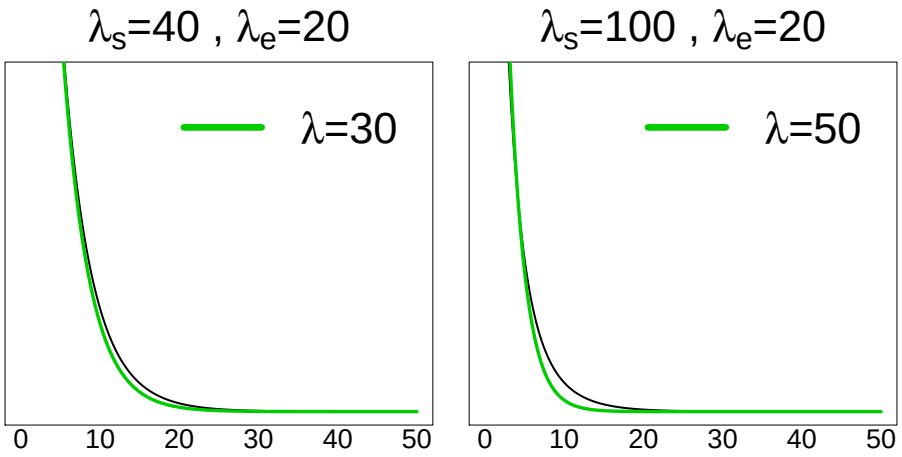



**Figure S1:** Continuous admixture simulated to occur between (left) 20 and 40 generations ago and (right) 20 and 100 generations ago. The black line shows the exponential decay curve predicted by Equation S5 with continuous migration ($\mu = 0.001$) between $[\lambda_s, \lambda_e]$ generations given in the title. The green line shows the best fit to the black line for a single instantaneous event curve from Equation S1, which has rate $\lambda$ given in the legend.

# S3 Chromosome painting, mixture modelling of admixing populations, and pairwise coancestry curves

In this section we describe how we use the theory developed in Note S2 in practice. Specifically, we first describe how we produce a "painting" of each genome in a population and then use this to (i) model the haplotypes in both this group, and in source groups that may have admixed to form this group, as a mixture of those in sampled groups, (ii) "clean" the painting to produce weights for each sampled group contributing to this mixture at each point in the genome, and (iii) study properties of these weights at pairs of positions at different distances $g$ apart in the genome, under different possible admixture models or when no admixture has occurred. In Note S4, we describe how we fit and test models based on our painting, so as to characterise admixture events, using the theory developed in this section.

## S3.1 Chromosome "painting" to make "copying vectors"

We use a model originally introduced by Li and Stephens (42), and since updated and implemented in the package CHROMOPAINTER (http://www.paintmychromosomes.com/) (8), to "paint" an individual's genome conditional on other sampled haplotypes. This procedure is described in detail in Appendix A. In human data, this painting produced by the Li and Stephens approach typically copies segments of each individual's genome from all other sampled individuals (Figure 1A-B of main text; (8)), even those in populations more distantly related to that which the individual being painted belongs. Thus, although we do see an enrichment of copying among closely related groups, this is subtle, or "noisy". This "noise" is in fact expected under plausible population genetic models of human evolution, since most human genetic variation is shared (43) and thus predates population splits, meaning that the relationships among haplotypes also frequently predate these splits. As sample sizes greatly increase in future, more recent relatives to a typical haplotype will be found, and such recent relatives are more likely to come from closer groups, suggesting improvements in the raw painting. Nevertheless, it is still necessary to take account of such noise in analyzing the history of a population, and we endeavour to do this by characterizing haplotypes of a group as a mixture of those of other

sampled groups, modeling the noise expected for such a mixture.

For the remainder of this supplement, we will make the following assumptions. We assume that individuals in a dataset are associated with population "labels" (e.g. sampled HGDP groups or fineSTRUCTURE-identified populations), some of which have been modified based on results of the clustering algorithm fineSTRUCTURE (8) (see Table S10 and Note S6.2). The initial chromosome painting results in a haploid genome sequence being decomposed into a series of chunks, each copied from the haplotypes of other members of the dataset (Figure 1 of main text). By associating these "donors" with their population label, we may regard each chunk as "coloured" by the label of the corresponding haplotype copied from, and thus the entire haploid chromosome is painted with a series of different colours, with as many colours as potential donor labels. Thus we define populations as vectors, termed "copying vectors" in the remainder of this supplement, representing the proportion of the genome an individual from the population copies from individuals of every other population on average.

The central idea of our approach is to use only this colouring, along the genome, which we can regard as a (detailed) summary of the haplotypic structure of a particular genome, to learn details of the admixture history of the population.

In general, suppose within a population, we paint a single chromosome of interest using a total of $K$ labeled groups. At any position in the genome, define $f_i, i \in [1, ..., K]$ to be the probability that position lies within a segment with a donor from population $i$, i.e. in a segment of the $i$th colour. We assume that the population is homogenous, so that $f_i$ does not differ among individuals within a group, but consider robustness to this assumption in simulations (Note S5). Note also that in general, other members of the same population as the chromosome of interest may also be potential donors, i.e. the group from which the chromosome of interest is painted may be one of the $i$s: we can "self-copy". We view the $f_i$s as one way of characterizing the population, in terms of its relationship with other groups. We will denote the $K$-vector of the $f_i$s as $f$.

More generally, we paint members of all sampled populations in exactly the same way, described in Note S4.1.1, using a compositionally identical set of possible "donor" chromosomes (not exactly identical only in that an individual will sometimes be replaced by another individual with the same population label). We define $f_i^l$ to be the average proportion of the genome painted using group $i$, for members of any group $l$, with corresponding $K$-vector $f^l$.

Now suppose that our population is admixed, between a total of $L$ sources labeled $S_1, S_2, ..., S_L$. In this setting, our chromosome of interest consists of unobserved chunks of these $L$ ancestries. We make the following assumptions regarding the relationship between our chromosome painting and these ancestries, whose effect we later test via simulation and/or robustness checks in our real data analysis:

1. Within an ancestry segment drawn from population $S_v$, the probability a randomly chosen position is painted using population $i$ is denoted by $f_i^{S_v}$, for $i \in [1, ..., K], v \in [1, ..., L]$, giving a $K$-vector $f^{S_v}$. For example for an admixing population $A$ we have a vector $f^A$. Note unless we sample the admixing groups, these quantities are not directly observable and we view them as characterising the admixing groups, so inferring these vectors, along with the proportions of ancestry from each source, is one of our key aims.

2. We assume these underlying copying probabilities are identical across the genome, so do not e.g. differ among chromosomes. Thus, we assume data quality and informativeness, averaged across individuals, and our ability to account for linkage disequilbrium (LD), do not vary strongly across the genome.

3. Copying chunks are assumed independent of one another within an ancestry segment, provided they are distinct, and separated by some minimum distance in the genetic map

– we use 1 centimorgan (cM), corresponding to around one megabase in humans on average, in this work. Thus we are assuming that at the 1cM distance within admixture chunks, LD breakdown is powerful enough to ensure independent relationships with other sample members.

4. Genetic drift specific to the population of interest can be modelled as an increased rate of self-copying (an excess of haplotypes shared with other members of the same group).

## S3.2 Using "copying vectors" to describe groups as mixtures of sampled populations

If a population is admixed, its haplotypes are descended from a mixture of those in the admixing source groups. Characterising these unknown groups genetically, using only the genomes of the admixed individuals, is a very complex and unsolved problem. To approximate this characterisation, we model haplotypes within each contributing group to an admixture event (e.g. from a specific region) as some unknown linear mixture of those in sampled populations, and infer details of this mixture. In doing this, we account for the noise discussed in Note S3.1 that is due to the painting itself. In the very simple setting where the group of sampled individuals are formed by a simple admixture of other sampled populations, clearly each source is trivially representable this way, and such a representation will reveal the admixing populations.

In practice, we believe this will rarely be the case. Frequently the groups mixing to form a present day admixed population may not be sampled, may have since drifted genetically, or may even be extinct. In this case, the hope is that the groups contributing in the mixture will be those sampled groups most closely genetically related to the true admixing population, and so offer information on this population – in Note S5 we test this directly via simulation. Allowing a mixture representation provides additional flexibility in better modelling such unsampled sources, because multiple sampled groups might jointly better capture similar haplotypes to those of the true source than any single group does. If the other sampled groups are themselves strongly admixed, this will add noise, but we anticipate the copying process will reduce this by only copying relevant genomic segments. Critically, at least in theory this should not lead to false admixture inference. Again we evaluate the effect of this issue in practice in Note S5.

The admixed population is in this setting a mixture of mixtures, and so can themselves be thought of as a mixture of sampled populations. In our approach we first fit this mixture directly, and later aim to deconstruct it in terms of contributing admixing groups. We find that the mixture representation typically allows an extremely accurate match to the distribution of colours in most groups (coefficient of determination $R^2 > 0.993$ in 92 of our 95 populations, using on average 20 groups with mixing coefficient above 0.001), suggesting this approach captures key features of genetic variation in real data. We discuss the performance of this approach in practice (via simulation based on real human populations) in Note S5.

We will illustrate the mixture representation in practice using the example setting of a single time and admixture event between two populations labeled $A$ and $B$. In a population with this ancestry, we suppose that a proportion $\alpha$ of the genome is drawn from population $A$, and a proportion $1-\alpha$ from population $B$. Then the probability a position in the genome is painted using population $i$ is given by

$$f_i = \alpha f_i^A + (1-\alpha) f_i^B.$$

(Note that $\alpha$ may differ from the proportion of chunks copied from population $A$, if chunks differ in size on average between the admixing groups.) Note that this quantity can be directly estimated from the painting alone, so in a large genome $f_i, 1 \leq k \leq K$ is known with considerable accuracy. In our data, a genome contains on average ≈20,000 copying chunks

across labeled populations in the "full analysis" described in Note S6.3, with an average size of ≈0.3-0.4cM. As before let $f$ denote the $K$-vector formed by these probabilities. Similarly, we have vectors $f^A, f^B$ corresponding to the populations contributing to the admixture event. Then we have:

$$f = \alpha f^A + (1-\alpha) f^B.$$

We suppose that for $1 \leq l \leq K$, the $K$-vector $f^l$ gives the average copying vector for a member of sampled population $l$, painted using the exact same $K$ sets of potential donors as used to paint the population of interest. Thus $f^l$ captures the "noise" expected in painting for an ancestry segment from population $l$. In practice we can estimate $f^l$ highly accurately by performing such a painting of all individuals in this population. The key step in our approach is that we model the vectors for each of the two contributing source groups to the admixture event as a mixture of the form:

$$f^A = \sum_{l=1}^{K} \gamma_l f^l, \qquad f^B = \sum_{l=1}^{K} \zeta_l f^l,$$

where $\gamma_l \geq 0$ and $\zeta_l \geq 0$ equal sampled population $l$'s mixing coefficients for source groups $A$ and $B$, respectively, which are then automatically constrained to sum to 1. Let $\vec{\gamma}$ and $\vec{\zeta}$ denote the two $K$-vectors containing the sets of these mixing coefficients.

This induces $f$ to be a mixture of the same type. Specifically, we have mixture coefficients $\beta_1, \beta_2, ..., \beta_K$ corresponding to the mixing coefficients for sampled populations $1, 2, ..., K$ and

$$f = \sum_{l=1}^{K} \beta_l f^l, \tag{S6}$$

where these overall mixture coefficients are given by:

$$\beta_l = \alpha\gamma_l + (1-\alpha)\zeta_l.$$

The situation in a setting with multiple groups is similar; we have one vector of mixing coefficients for each admixing group and one $\alpha$ for each event, and Equation S6 applies in general. We think of only a subset of the $\beta_l$s – those groups truly related to the population being considered – as non-zero, simplifying the problem. For example if our population is an admixture of two exact matches to sampled groups, only the corresponding two $\beta_l$ terms will be non-zero, with one equal to $\alpha$ and the other to $1-\alpha$ in our single admixture event setting. (The $\beta_l$s may be estimated as described in Note S4.2, so can be regarded as approximately known – we assume the vectors $f^l$ are linearly independent, which is expected to be the case provided each group has undergone a small amount of genetic drift relative to the others.)

There is one additional important detail that we incorporate into our approach, which is that genetic drift specific to the population being considered is often likely to have occurred, because our sampled groups are drifted relative to one another and to any true source populations. As stated above, this is assumed to result in an excess of "self-copying" with a drifted group copying from itself more than expected according to any strict mixture representation. To account for this, we conceptually introduce a "sampled" group that copies *only* from the population considered. (Mathematically, provided – as we always observed in practice – drift is non-negative, this is equivalent to ignoring, or normalizing to zero, the fraction of the genome self-copied, in fitting the other mixture coefficients.) The phenomenon of potentially recent drift also means that information on self-copying parts of the genome does not provide directly useful information about e.g. admixture dates, and so in order to later infer details of the admixture history of a group, we perform a second painting where self-copying is not allowed.

The $\vec{\gamma}$, $\vec{\zeta}$, and $\alpha$ terms thus constructed help to characterize the groups involved in the admixture event, since they give both the makeup of the admixing groups (as a mixture of sampled groups) and the admixture proportion $\alpha$. Note also that together, these define the vector $f$ of painting probabilities, which we can estimate from data, so we can seek least-squared estimates of $\vec{\gamma}$, $\vec{\zeta}$, and $\alpha$ (Note S4). For this estimation to be possible, given the $2K + 1$ or more parameters to be estimated, we utilise additional information, which we obtain by also minimising squared error to additional measures based on "coancestry curves" detailed in the following sections.

## S3.3 Generating a "cleaned painting" by defining weight vectors for each copying chunk

The representation of haplotypes in a population as a mixture of those in other sampled groups with coefficient vector $\vec{\beta}$ effectively reduces dimension, because in practice most of the coefficients $\beta_m \geq 0$ are normally either zero or of small size. To translate this into our painting, we produce a "cleaned" (reweighted) painting by constructing a weight matrix $W_{mi}$ that gives the probability that a chunk has ancestry from group $m$ in the mixture, given that it is copied from group $i$. By Bayes' formula this conditional probability is proportional to the prior probability $\beta_m$ of group $m$ multiplied by the probability of copying from group $i$ given a haplotype comes from group $m$, given by $f_i^l$ from Note S3.1. Unconditionally this defines the weight matrix:

$$W_{mi} = \frac{\beta_m f_i^m}{\sum_{j=1}^{K} \beta_j f_i^j}. \tag{S7}$$

The set of values $W_{mi}, 1 \leq m, i \leq K$ form a fixed square matrix $W$, which we can interpret as a weighting function allowing for noise in the painting. In practice, we find this matrix is typically sparse with multiple zero rows, due to non-contributing populations.

We apply this weighting function to each copying chunk, to obtain a vector of weights for every position $s$ in the genome. Specifically, $s$ lies within some chunk, and this chunk is copied from a member of some population $i(s)$ corresponding to a particular column of $W$. Then we define the cleaned painting at $s$ as the vector of entries $Q_{ms}$, $1 \leq m \leq K$ where $Q_{ms} = W_{mi(s)}$. As shown in Figure 1B of the main text, this approach reduces noise, but we note it only uses information from a single chunk, and not e.g. additional information that might be present from correlated ancestry along the genome. This chunk-by-chunk reweighting is a key property for the theoretical underpinnings of our method, but means our cleaned painting is **not designed for inferring local ancestry**.

## S3.4 Generating coancestry curves using weights and the cleaned painting

In practice, we cannot usually observe admixture chunks but must infer their properties indirectly, using (in our approach) chromosome painting (see Note S3.1). As we described in Note S3.3 we generate a cleaned painting that attaches a weight vector $Q_{1s}, Q_{2s}, Q_{3s}, ..., Q_{Ks}$ to each position $s$ in the genome, corresponding to the probability of haplotype sharing with each of $K$ sampled populations at $s$. Because these weights are based on an initial painting of each individual in terms of the same $K$ groups, the assumptions listed in Note S3.1 of homogeneity across the genome of the distribution of the chunk copying probabilities and independence of chunks separated by some minimum distance (e.g. 1cM) in the genetic map in this initial painting, conditional on underlying ancestry, immediately mean the same properties hold for the weight vector.

Suppose we have unobserved true admixing source groups labelled $S_1, S_2, ..., S_L$. Given the homogeneity across the genome, conditional on the source group being $S_v$ we can define *expected* weights $Q_m^{S_v} = E(Q_{ms} \mid \text{ancestry } S_v)$ for $1 \leq m \leq K$ and $1 \leq v \leq L$, valid for any position $s$ in the genome. Note that these cannot be directly estimated from the data without knowledge of the true ancestry, and that $L$ and $K$ are not normally the same. For example for a simple admixture event between populations $A$ and $B$, we have mean weights $Q_1^A, Q_2^A, ..., Q_K^A$ and $Q_1^B, Q_2^B, ..., Q_K^B$.

The independence of the weights means at any two positions $l$ and $r$, a sufficiently large genetic distance $g$ apart, that conditional on the underlying ancestry, the "cleaned painting" weights at these positions, $Q_{ml}$ and $Q_{nr}$, are independent for pairs of possible populations $1 \leq m, n \leq K$.

Although in this work we use the approach to construct weights given in Note S3.3, other authors have suggested differing schemes, based on single markers (4). We note that the subsequent theory applies for a variety of possible weighting schemes, provided they produce weights uniform on average across the genome, and that satisfy the independence property above for suitably distant pairs of positions. In particular, it applies when we do not allow a group to "self-copy" (Note S3.2), as is the case for the painting we use to generate the following "coancestry curves" in practice. As discussed further below, it also applies if our weights are "inaccurate" in terms of how well the mixture representation matches the observed haplotype patterns.

**Expected weight products:** Admixture in the history of our sample induces correlation in our weights at different distances because of correlations in the underlying ancestry, and this is the key property we wish to exploit in inferring such events.

One natural measure of this correlation is the product of expected weights for positions $l$ and $r$ separated by distance $g$. For any $1 \leq m, n \leq K$ this is given by, using notation as in Note S2.1:

$$\begin{aligned} E(Q_{ml}Q_{nr}; g) &= \sum_{v=1}^{L}\sum_{w=1}^{L} E(Q_{ml}Q_{nr} \mid \text{endpoint ancestries } S_v,S_w) p_{S_vS_w}(g) \\ &= \sum_{v=1}^{L}\sum_{w=1}^{L} Q_m^{S_v} Q_n^{S_w} p_{S_vS_w}(g) \\ &= Q_mQ_n + \sum_{v=1}^{L}\sum_{w=1}^{L} Q_m^{S_v} Q_n^{S_w} [p_{S_vS_w}(g) - p_{S_v}p_{S_w}], \end{aligned} \tag{S8}$$

where $Q_m \equiv \sum_{v=1}^{L} Q_m^{S_v} p_{S_v}$ is the average weight for population $m$, and the second line follows from the independence assumption.

We simply normalise these curves by these average weights to form "coancestry curves" which are central in our admixture inference procedure:

$$\Psi(Q_{ml}Q_{nr}; g) \equiv \frac{E(Q_{ml}Q_{nr}; g)}{Q_mQ_n} = 1 + \frac{\sum_{v=1}^{L}\sum_{w=1}^{L} Q_m^{S_v} Q_n^{S_w} [p_{S_vS_w}(g) - p_{S_v}p_{S_w}]}{Q_mQ_n}. \tag{S9}$$

From Equation S8, the dependence of these coancestry curves on the distance $g$ between endpoints depends only on the expected weights conditional on the true underlying populations at each endpoint, and on the correlation in underlying ancestry at distance $g$, $p_{S_vS_w}(g)$. We can apply Equations S8 and S9 to consider various possible admixture histories, as studied in Note S2 for the underlying ancestry, with the key difference being that while this underlying ancestry is not observable, the coancestry curves based on weights can be directly estimated from properties of these weights along the genome.

## S3.5 Weight-based coancestry curves for a single admixture event

Using the same notation as Note S2, we have from Equations S1, that for a simple admixture event at a single time between two groups $A$ and $B$, and with admixture fraction $\alpha$:

$$\begin{aligned} p_{AA}(g) - p_A p_A &= \alpha(1-\alpha)\exp^{-g\lambda} \\ p_{AB}(g) - p_A p_B &= -\alpha(1-\alpha)\exp^{-g\lambda} \\ p_{BB}(g) - p_B p_B &= \alpha(1-\alpha)\exp^{-g\lambda}, \end{aligned}$$

and applying to Equation S8, we find after simplification:

$$\begin{aligned} E(Q_{ml}Q_{nr};g) &= Q_m Q_n + \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]\exp^{-g\lambda} \\ &= Q_m Q_n + H_{mn}\exp^{-g\lambda} \\ &= Q_m Q_n + D_m D_n \exp^{-g\lambda} \end{aligned} \tag{S10}$$

Here we define $H_{mn} \equiv \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]$, and for $1 \leq m \leq K$ we have $D_m \equiv \sqrt{\alpha(1-\alpha)}[Q_m^B - Q_m^A]$. The (normalised) coancestry curve is then:

$$\Psi(Q_{ml}Q_{nr};g) \equiv \frac{E(Q_{ml}Q_{nr};g)}{Q_m Q_n} = 1 + \frac{D_m D_n}{Q_m Q_n}\exp^{-g\lambda} = 1 + \delta_{mn}\exp^{-g\lambda}. \tag{S11}$$

There are a number of important implications from this formula that apply across possible weighting schemes.

1. As $g$ varies, the right hand side describes an exponential decay curve, with rate the (true) time since admixture $\lambda$, allowing estimation of admixture time directly from the curve. This approach will work for any set of weights satisfying the assumptions, provided at least one of the terms $\alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]$ is non-zero. Given some admixture, we have $0 < \alpha < 1$, and so this is guaranteed (taking $m = n$) provided $Q_m^B - Q_m^A \neq 0$ for some $m$. Thus, provided the expected value of at least one weight differs depending on the true underlying ancestry, we may estimate admixture time.

2. The coefficients $H_{mn}$ relating to the $m$,$n$th curve form a matrix $H$ of coefficients. Taking a vector $\vec{D}$ with elements $D_m$, we have $H = \vec{D}^T\vec{D}$ and thus $H$ is of rank 1 (equivalently, its eigendecomposition includes only one non-zero eigenvalue). This fact can be used to distinguish single admixture events from more complex events, given a sufficiently rich set of weights (Note S3.6).

3. Given real data we may calculate weights for both chromosomes an individual carries (dealing with unknown haplotypic phase where necessary) in order to empirically estimate $E(Q_{ml}Q_{nr};g)$, $Q_m$ and $Q_n$ for each $m$ and $n$, and as a function of $g$. This allows us to estimate the coancestry curves on the right hand side empirically, given an estimated recombination map (e.g. (44),(45),(46)), and to evaluate whether these curves fit the simple form predicted by Equation S11. To do this, we average across all pairs of positions, chromosomes, and individuals in the group of interest, potentially yielding highly accurate curve estimates. Given the curves for all $m$ and $n$ (assuming a single admixture date), it is then possible to fit (Notes S4.4-S4.5) (i) the decay rate $\lambda$ based on least squares estimation, and at the same time (ii) the matrix $H$, to help in characterising the admixture event. If no admixture occurs, we would expect $E(Q_{ml}Q_{nr};g)$ to show no clear pattern with genetic distance (under our modeling assumptions), and so the fitted curve should have little predictive power relative to a constant value. Testing whether the matrix $H$ has a single dominant eigenvalue aids in distinguishing single and more complex admixture events, particularly where these occur simultaneously or near-simultaneously. If a single eigenvalue dominates, the corresponding eigenvector forms an estimate of $\vec{D}$ after rescaling by the square root of this eigenvalue.

4. Each curve corresponds to a comparison for a pair of populations involved in the mixture inferred for the (potentially) admixed groups, and can be labeled by these groups, e.g. "Balochi-Mandenka" or "Mandenka-Mandenka" (Figure 1C in the main text).

5. Any two events yielding the same values $D_m$ will yield exactly the same collection of curves, so the admixture fraction $\alpha$ is not always identifiable from the curves in general (but may be in many realistic cases – see below). The value of $D_m$ is proportional to the difference in average weight given to population $m$ between segments of the genome coming from admixing group $B$, and segments coming from admixing group $A$. If population $m$ is more closely related to admixing group $B$, we would expect this quantity to be positive, so its value indicates which groups are more closely related to which contributing admixing populations. For example, in simulations admixing Yorubans (population $B$) and Brahui (population $A$) and then analysing the admixed group treating these groups as unsampled, one of the inferred populations in the mixture is always Mandenka (Figure 1 of the main text, Note S5). This group is more closely related to the Yoruba than the Brahui population. Thus the corresponding value of $D_m$ for Mandenka is positive, because the expected weight for the Mandenka component in the mixture is greater given true underlying Yoruba ancestry than Brahui ancestry. We can then regard a high Mandenka weight as a (noisy) surrogate for being in a Yoruba segment. Similarly, a high Balochi weight is a surrogate for being in a Brahui segment, and the Balochi obtain a negative $D_m$. Finally, at very short distances $g$ it is highly unlikely one end of an interval of length $g$ is in a Yoruba segment and the other end is in a Brahui segment. Correspondingly, we would expect the expected weight product for the probability of Mandenka and Balochi to be reduced at short distances relative to long. This is both seen in the corresponding curve (Figure 1C) and predicted by Equation S11, given the negative sign of $D_m D_n$ in this case. For "diagonal" curves $m = n$, e.g. the Mandenka-Mandenka weight curve, we will always see an increase at short distances, because there is an increased chance both endpoints come from whichever admixing group population $m$ is most closely related to (Figure 1C of main text).

6. The maximum possible absolute value of the "coefficient" term $H_{mn} \equiv \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]$ is $\alpha(1-\alpha)$, and this is achieved only if the absolute value of the weights for populations $A$ and $B$ differ by 1, i.e. if the weights for contributing groups $m$ and $n$ are perfect surrogates for true ancestry. In this setting (which we never saw in simulations, or real human data, due to small differences among human groups) the corresponding curve matches $p_{S_v S_w}$ for some $S_v$ and $S_w$, so this setting is (obviously) equivalent to having full ancestry information along the genome. Thus, the effect of noise in the painting in more realistic settings is that we will see a far smaller difference for the plotted curves between the values at $g = 0$ and $g = \infty$ than if the ancestry were truly known. In general this reduced signal results in more dating uncertainty for a given level of noise (i.e. amount of data), so unsurprisingly closer population surrogates within the sample improve the performance of our approach. Even given close surrogates, strong similarity among human groups means a single linkage disequilbrium (LD) chunk, and hence our method, will not typically strongly determine ancestry. In practice, we typically predict by modeling, and observe in practice, tiny fitted values for the coefficient term, in some cases well below 0.1% of the asymptotic value. Nevertheless, the huge volume of data available from the entire genome allow correct inference of admixture history details even in many such cases (Note S5).

7. In practice, our inference procedure for estimating the weighting function used to generate the values of $Q_{ml}, Q_{nr}$ is obviously subject to errors given real data, and our mixture rep-

resentation cannot precisely represent the true admixing groups. However, our procedure always results in some weighting function, which is a linear transformation of the observed painting. The derivation of Equation S11 in fact in no way depends on the particular form of this transformation, and so in general an equation of the same form still holds with, critically, the same exponential rate parameter. This means that estimation of the admixture time from the curve will still yield good results, even when the procedure to determine admixing groups (detailed below) is subject to error. Further, estimation of this time is robust to arbitrarily inaccurate inferred mixtures, or uninformative sampled donor groups. (On the other hand, power to detect admixture may be lower in such cases.)

## S3.6 Weight-based coancestry curves for a double admixture event

Next, applying Equation S8 using equations of the form of Equations S2, we find in the previously analysed (Note S2.3) setting of admixture involving two admixture events, using the same notation as in that note:

$$E(Q_{ml}Q_{nr};g) = Q_mQ_n + \alpha_3(1-\alpha_3)[\alpha_1Q_m^A + \alpha_2Q_m^B - Q_m^C][\alpha_1Q_n^A + \alpha_2Q_n^B - Q_n^C]\exp^{-g\lambda_2} \\ +(1-\alpha_3)\alpha_1\alpha_2[Q_m^B - Q_m^A][Q_n^B - Q_n^A]\exp^{-g\lambda_1}. \tag{S12}$$

The coancestry curves simply divide by the constant first expectation term on the right hand side. In general, if the admixture times are different, we note that this predicts the observed coancestry curves $E(Q_{ml}Q_{nr};g)$ are mixture of exponential curves, with rate parameters $\lambda_1$ and $\lambda_2$ respectively. Thus, we can distinguish a single admixture time from two admixture times by comparing the fit of a mixture of exponentials to that of a single exponential (Notes S4.5, S4.6, S4.8) for the observed coancestry curves. Given this, we can form two coefficient matrices $H_1$ and $H_2$ corresponding to the two admixture decay rates. Using the same argument as in the single admixture event case, it is immediately clear that each of these matrices will be of rank 1; they can be interpreted as corresponding to their respective admixture events.

An important special case comes where two admixture events occur simultaneously (3-way admixture). In this case, using the notation of Equations S4 we find after simplification:

$$E(Q_{ml}Q_{nr};g) = Q_mQ_n + \Big(\beta_A\beta_B[Q_m^B - Q_m^A][Q_n^B - Q_n^A] \\ +\beta_A\beta_C[Q_m^C - Q_m^A][Q_n^C - Q_n^A] + \beta_B\beta_C[Q_m^C - Q_m^B][Q_n^C - Q_n^B]\Big)\exp^{-g\lambda}. \tag{S13}$$

Since there is only a single admixture time, we can identify this time (as in the single admixture event) by fitting a single rate. Also as previously, we can form a single matrix $H$ of coefficients multiplying the exponential rate:

$$H_{mn} = \beta_A\beta_B[Q_m^B - Q_m^A][Q_n^B - Q_n^A] + \beta_A\beta_C[Q_m^C - Q_m^A][Q_n^C - Q_n^A] + \beta_B\beta_C[Q_m^C - Q_m^B][Q_n^C - Q_n^B].$$

However, in this setting $H$ has a more complex form, allowing (in principle) us to distinguish multi-way admixture from simple 2-way admixture. Specifically, if (analogously to the 2-way admixture case) we define two vectors $\vec{V}$,$\vec{W}$, with the following $m$th elements:

$$V_m = \sqrt{\beta_A\beta_B}[Q_m^B - Q_m^A] \\ W_m = \sqrt{\beta_B\beta_C}[Q_m^C - Q_m^B].$$

We will assume $\vec{V}$, $\vec{W}$ are not collinear; that is, the average weight differences conditional on underlying ancestry point in different directions in $K$-space, meaning the average weight vectors have some ability to distinguish among all three groups.

Then noting:

$$X_m = \sqrt{\beta_A \beta_C}[Q_m^C - Q_m^A] = \sqrt{\beta_A/\beta_B} W_m + \sqrt{\beta_C/\beta_B} V_m,$$

we can immediately write:

$$H = \vec{V}^T \vec{V} + \vec{W}^T \vec{W} + \vec{X}^T \vec{X}$$

and substituting for $\vec{X}$:

$$H = \big(1 + \frac{\beta_A}{\beta_B}\big)\vec{V}^T \vec{V} + \big(1 + \frac{\beta_C}{\beta_B}\big)\vec{W}^T \vec{W} + 2\frac{\sqrt{\beta_A \beta_C}}{\beta_B}\vec{V}^T \vec{W}.$$

Thus $H$ has rank 2, since all its rows (or columns) are linear combinations of the two vectors $\vec{V}$ and $\vec{W}$. It is straightforward to generalize to the case of simultaneous multi-way admixture involving $G$ groups. In this general case, $H$ has rank $G-1$. Since $H$ can be estimated directly from the data, in settings where there is no evidence of multiple rates in the exponential curve fit, so that the data are consistent with some number of populations mixing simultaneously, we can still (in principle) detect a signal of multi-way admixture where it occurs, based on the number of "large" eigenvalues of the fitted coefficient matrix $H$ (Note S4.9). In practice, we focus on the lower-dimension cases $G = 2$ or 3.

## S3.7 Weight-based coancestry curves for continuous admixture

Finally, applying Equation S8 using equations of the form of Equation S5, we find in the previously analysed (Note S2.4) setting of admixture involving continuous migration of a single group into another:

$$E(Q_{ml}Q_{nr}; g) = Q_m Q_n + \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A] \sum_{j=\lambda_e}^{\lambda_s} w_j \exp^{-gj}.$$

That is, in this setting and exactly as for the unobserved underlying ancestry, the curves $E(Q_{ml}Q_{nr}; g)$ and hence, following normalisation, the coancestry curves are mixtures over weighted individual exponential curves with decay rates, as $g$ increases, in the range $[\lambda_e, \lambda_s]$. In fact, it can be shown that an equation of the same form, with appropriately redefined weights $w_j$, holds in the more general setting of variable admixture rates through time, where migrants always come from a single group. A particular special case of interest is admixture at two times, where only two $w_j$ terms, corresponding to these times, are non-zero. Once again, we have a mixture of decay coefficients, but in this case all such coefficients are proportional and determined by $H_{mn} = \alpha(1-\alpha)[Q_m^B - Q_m^A][Q_n^B - Q_n^A]$, which will be greater when the weights are effective at distinguishing the underlying groups, or where $\alpha$, the current fraction of ancestry from the migrant group, is close to $\frac{1}{2}$.

As discussed previously (Note S2.4), we believe it likely to be extremely difficult in practice to distinguish multiple "pulse-like" admixture events from a more continuous occurrence of admixture, due to the difficulty in resolving such exponential mixtures. However, we see many cases in real data, and in simulations, where it is possible to distinguish a single admixture time from multiple admixture times. Note that in the case considered here, where the same group is involved at each time, the above shows that the coefficient matrix $H$ for each event is the same up to a constant (and of rank 1). Thus we can distinguish this case from that considered above with different populations involved at each time, where we see two distinct coefficient matrices $H_1$ and $H_2$, again each of rank 1. Thus, it is possible in general to distinguish multi-time admixture events involving the same groups from those involving different groups, by comparing the first eigenvector of the coefficient matrices corresponding to each time.

# S4 Fitting the admixture event model to identify admixture times and admixing groups in practice

In this Note, we describe in detail how we infer admixture events, their dates and proportions, and the genetic make-up of the source groups involved, for the analyses presented in this paper. We concentrate first on the procedure for testing whether admixture is present, and inferring a single date for a single simple admixture event. In Notes S4.8 and S4.9 we describe how our inference varies from this procedure when investigating complex events. We will refer to results derived in Note S3 extensively, where the assumptions made are described.

## S4.1 Protocol for chromosome painting

### S4.1.1 chromosome painting to estimate copying vectors

Using the notation of Note S3, we refer to the "copying vector" $f^i$, of a sampled population $i$ as a vector with elements containing the proportion of DNA that population $i$ copies from each other population in the dataset including its own under the CHROMOPAINTER (8) model (see Note S3.1 and Appendix A). These copying vectors describe how populations relate to one another in terms of the relative time to a common shared ancestor, subsequent recent admixture, and population-specific drift.

To generate copying vectors for this analysis, unless otherwise noted we perform a "leave-one-out" procedure where each individual from a given population $k$ is allowed to copy from every other individual with the same population label and the first $n_l - 1$ individuals from each donor population $l \neq k \in [1, ..., K]$, with $n_l$ the number of individuals with population label $l$. We aim to paint each individual using all other samples in order to learn about ancestral relationships. Then there are $n_l$ samples to copy from each population $l \neq k \in [1, ..., K]$, while only $n_k - 1$ samples to copy from their own population (as they cannot be used to paint themselves). To avoid this reduction by 1 causing problems later, we instead removed one individual from each of the other populations $l \neq k \in [1, ..., K]$ when painting. Thus all individuals in the dataset copy from the same number of individuals from each labeled population.

For each individual, we ran CHROMOPAINTER (8) with 10 Expectation-Maximisation (E-M) steps to jointly estimate the program's parameters $N_e$ and $\theta$ (see Appendix A.3-A.4), repeating this separately for chromosomes 1, 4, 10, 15 and weight-averaging (using centimorgan sizes) the $N_e$ and $\theta$ from the final E-M step across the four chromosomes. We then averaged these $N_e$ and $\theta$ estimates across all individuals. Finally, using these individual-averaged values of $N_e$ and $\theta$, we re-ran CHROMOPAINTER one more iteration for each individual to estimate $f^i$, the total proportion of genome-wide DNA copied from each labeled population, using the procedure described in Appendix A.5. Note that we therefore use the same values of $N_e$ and $\theta$ in CHROMOPAINTER when generating each individual's final copying vector, so that copying vectors across individuals are directly comparable. (The strict definitions of $N_e$ and $\theta$ are provided in Appendix A, but the rough intuition is that $N_e$ helps determine the average size of donor segments copied in the painting, and $\theta$ captures the proportion of SNP mismatches between donor and recipient.)

### S4.1.2 chromosome painting to generate coancestry curves

To infer the probability that an individual copies from any other individual at a specific locus or SNP, we form the genome of each individual as a mosaic of the genomes of other individuals in the dataset, i.e. we "paint" the chromosomes of each individual as described in Note S3 and Appendix A.6. A property of our painting algorithm is that it seeks to find each individual's

closest relative in the dataset at each genome location. Two highly related individuals will therefore copy a large proportion of their DNA from one another under our model. This extends to populations as well, in that due to genetic drift, individuals will often paint a large proportion of their DNA using members with the same population label. We therefore disallow copying from individuals with the same population label when painting chromosomes (Note S3.2), in order to avoid masking the signal of admixture common to all the population's samples.

For each individual in a given population, we jointly estimate $N_e$ and $\theta$ using 10 E-M steps of CHROMOPAINTER, repeating this separately for chromosomes 1, 4, 10, 15 and again weight-averaging the $N_e$ and $\theta$ from the final E-M step across these four chromosomes. We then average these $N_e$ and $\theta$ estimates across all individuals in the population and rerun CHROMOPAINTER one more iteration for each individual using these fixed values, generating 10 painted chromosome samples for each haplotype. Note that in contrast to the copying vector protocol described in Note S4.1.1, each population uses its own specific $N_e$ and $\theta$ estimates when generating painting samples, reflecting the fact they are not allowed to self-copy in this component of the analysis.

## S4.2 Initial fitting of population haplotypes as a mixture of those of other groups

In Note S4.1.1 and Appendix A.5, we describe our protocol and formula, respectively, for calculating the "copying vector" for a population. Using analogous notation to Note S3.2, let $\hat{f}_i^k$ be the final estimated contribution from donor population $i$ averaged across all individuals in recipient population $k$. Let $\hat{f}^k \equiv \{\hat{f}_1^k, ..., \hat{f}_K^k\}$ be the observed "copying vector" of proportions for recipient population $k$, so that $\sum_{i=1}^K \hat{f}_i^k = 1.0$. We calculate the corresponding observed "copying vector" $\hat{f}^l$ for each population $l \in [1, ..., K]$ and view this as an estimate of the true underlying mean copying vector for that population.

We perform an initial estimate of the mixing coefficients that describe the copying vector of a putatively admixed population $k$ as a mixture of those of other populations. For all of our populations, we find that they copy more from themselves than any other population copies from them. We interpret this as evidence of drift, which can often be substantial. As described in Note S3.2, we account for this excess of self-copying by (conceptually) introducing a coefficient in our mixture model to capture it. Operationally, since this coefficient is not of direct interest, we can simply remove entries corresponding to copying from population $k$, which gives identical values in terms of admixture inference. I.e. we subtract out the proportion of genome that each population $1, ..., K$ copies from population $k$ under our copying model. In particular we set $\hat{f}_k^l = 0$ and rescale such that $\sum_{i=1}^K \hat{f}_i^l = 1.0$ for all $l \in [1, ...K]$. We let $\hat{f}^{*l}$ represent the vector for population $l$ rescaled in this manner, with each element $i \in [1, ...K]$ represented as $\hat{f}_i^{*l}$ (note again that $\hat{f}_k^{*l} = 0$).

Allowing for errors in Equation S6 of Note S3.2, we assume a standard linear model form for the relationship between $\hat{f}^{*k}$ and terms $\hat{f}^{*l}$ for $l \neq k \in [1, ..., K]$, i.e.:

$$\hat{f}^{*k} = \sum_{l \neq k}^{K} \beta_l^k \hat{f}^{*l} + \epsilon,$$

where $\epsilon$ is a vector of errors, which we seek to choose the $\beta$ terms to minimise using least-squares. Here $\beta_l^k$ is the coefficient for $\hat{f}^{*l}$ under the mixture model. We use the non-negative-least-squares "nnls" package in R to estimate the $\beta_l^k$s under the constraints that all $\beta_l^k \geq 0$ and $\sum_{l \neq k}^K \beta_l^k = 1.0$. We refer to the estimated coefficient for the $l^{\text{th}}$ population as $\hat{\beta}_l^k$. To avoid over-fitting, in practice we exclude all populations for which $\hat{\beta}_l^k \leq 0.001$ and re-scale so that

$\sum_{l\neq k}^{K}\hat{\beta}_l^k = 1.0$. We let $T^*$ denote the set containing all populations $l \neq k \in [1, ..., K]$, for which $\hat{\beta}_l^k$ (prior to rescaling) is $> 0.001$.

The $\hat{\beta}_l^k$s represent our initial estimates of "true" mixing coefficients that describe the recipient population's DNA as a linear combination of the donor populations' DNA, in a robust manner that cleans up the imperfect inference of our raw painting algorithm. In particular it identifies donor populations whose copying vectors as inferred by the painting algorithm match the copying vector of the recipient population as inferred by the painting algorithm. Suppose a truly "San" segment on average is inferred by our painting algorithm to copy 50% of their genome from other San individuals, 40% from individuals of other African populations, and various smaller percentages from other individuals in the dataset, representing considerable noise. If in this example population $k$ were genetically identical, so that the copying vector for recipient population $k$ was identical to this toy San copying vector, then $\hat{\beta}_{\text{San}}^k$ and $\hat{\beta}_k^{\text{San}}$ would both be 1.0, removing the noise due to the painting for $k$.

## S4.3 Generating observed "coancestry curves"

In this section we describe how we estimate the coancestry curves defined in Note S3.4 in practice, based on having constructed a "cleaned painting" (Note S3.3), where we use a weight matrix $W$ to reduce noise in an initial set of "painting samples".

In Note S4.1.2 and Appendix A.6, we describe our protocol and formula, respectively, for generating such "painting samples" for an individual from population $k$. Using analogous notation to Appendix A.6, let the $L$-vector $\vec{X}^a \equiv \{X_1^a, ..., X_L^a\}$ be painting sample $a$ for one haploid of an individual from population $k$, with $X_l^a$ listing the donor individual copied at SNP $l \in [1, ..., L]$. Here we use ten painting samples representing each haploid; so in total we have $\vec{X}^1, ..., \vec{X}^{20}$ painting samples across both haploids for a given individual.

Define a "chunk" to be a segment of contiguous SNPs copied from a single haplotype of some donor population according to $\vec{X}^a$. For every pairing of painting samples between and within the recipient individual's two haploids, we compare each chunk on one sample to each chunk on the other sample, tabulating the donor populations represented by each chunk and the genetic distance $g$ between the two chunks' midpoints. Specifically, let $l$ be the first SNP within a chunk of size $w_l$ (in centimorgans; cM) from painting sample $\vec{X}^a$ and $r$ be the first SNP within a chunk of size $w_r$ from $\vec{X}^b$, with the midpoints of the two chunks separated by genetic distance $g$. Furthermore, let $\chi_{a,b;g}$ be the set of all chunk pairs with midpoints separated by distance $g$ and with one chunk from $\vec{X}^a$ and the other from $\vec{X}^b$. We find for each pairing of donor populations $i, j \neq k \in [1, ..., K]$:

$$\Phi_k(i, j; g) \equiv \sum_{a=1}^{20}\sum_{b=1}^{20}\sum_{\chi_{a,b;g}} \tilde{w}_l\tilde{w}_r, \tag{S14}$$

with

$$\tilde{w}_l = \begin{cases} \min(w_l, 1.0) & \text{if } X_l^a = i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\tilde{w}_r = \begin{cases} \min(w_r, 1.0) & \text{if } X_r^b = j \\ 0 & \text{otherwise.} \end{cases}$$

Thus longer chunks (in terms of cM) contribute relatively more to $\Phi_k(i, j; g)$ in proportion to their size, but we cap the weight $\tilde{w}_l$ of any chunk $l$ to be 1.0. This is so that the relatively small number of chunks >1cM in size (which might indicate recent sharing Identity-by-Descent,

IBD, i.e. recent relatedness) do not have a strong effect on inference, so that our approach is not strongly informed by IBD information. In both our real data analysis and simulations, we also explored removing chunks larger than 1cM (of which there are often relatively few), and results were similar. Considering chunks on samples from different haploids within the same recipient individual fully accounts for phase "switch errors", a common source of error in phasing (47). In Note S4.3.1 below, we describe how we group counts of $\Phi_k(i,j;g)$ into different bins of $g$.

Let $n_T = |T^*|$, i.e. the number of donor populations contained in $T^*$, defined in Note S4.2, which describe the DNA of recipient population $k$ under our mixture model.

The curves $\Phi_k(i,j;g)$ formed as $g$ varies count the length of genome observing respective ancestries $i$ and $j$ a distance $g$ apart in the "raw" painting, while the coancestry curves defined in Note S3.4 instead utilise the "cleaned painting" (Note S3.3). However, we can obtain curves relating to this latter painting as simple combinations of the "raw" curves. Storing these "raw" curves makes it possible to efficiently generate coancestry curves, using the raw curves, across iterations of inferring the mixture components and weights. Specifically, we note that if two populations $m$ and $n$ are in the inferred mixture, chunks copied from groups $i$ and $j$ will contribute weights $W_{mi}$ and $W_{nj}$ respectively to these groups in the cleaned painting. Then the corresponding curves for the "cleaned" painting are:

$$\begin{aligned}\Phi^*_k(Q_{ml}Q_{nr};g) &\equiv \textstyle\sum_{i,j\neq k}^{K} W_{mi}W_{nj}\Phi_k(i,j;g) \\ &= \textstyle\sum_{i,j\neq k}^{K} \big[\{c_i^k\hat{\beta}_m^k\hat{f}_i^{*m}\}\{c_j^k\hat{\beta}_n^k\hat{f}_j^{*n}\}\Phi_k(i,j;g)\big],\end{aligned} \tag{S15}$$

with $c_i^k = [\sum_{h=1}^{n_T}\hat{\beta}_h^k\hat{f}_i^{*h}]^{-1}$. For each distance $g$, this gives an $n_T \times n_T$ matrix of values. Note that (S15) represents our empirical estimate of $E(Q_{ml}Q_{nr};g)$ defined in Equation S10 of Note S3.5, based on observed counts. To obtain estimated coancestry curves, we must normalise by an estimate of the product of average weights $Q_mQ_n$. For each individual in $k$, we thus calculate the expectation for (S15) marginally for the left and right chunk pairs. We calculate this expectation by summing over all possibilities of the other endpoint, accounting in the denominator for the length of genome examined, as:

$$\widetilde{\Phi}^*_k(Q_{ml}Q_{nr};g) \equiv \big[\sum_{h=1}^{n_T}\Phi^*_k(Q_{ml}Q_{hr};g)\big]\big[\sum_{h=1}^{n_T}\Phi^*_k(Q_{hl}Q_{nr};g)\big]/\big[\sum_{h,p}^{n_T}\Phi^*_k(Q_{hl}Q_{pr};g)\big]. \tag{S16}$$

For each distance $g$, this gives an additional $n_T \times n_T$ matrix. If the populations $m$ and $n$ do not indicate evidence of admixture in this recipient individual, then (S16) should be approximately equivalent to (S15) at every distance $g$ for all $m,n$ in $T^*$. Indeed as $g$ gets large, we expect this to eventually become true for every $m$ and $n$, as the chunk you start from should provide no information on the chunk you end up at beyond genome-averaged probabilities of copying from $m$ and $n$, for example if the two chunks are on separate chromosomes (though in practice we only consider chunks on the same chromosome).

For each genetic distance $g$, we can now average left/right endpoint status and perform the normalisation to make a new symmetric matrix $\widehat{\Psi}_k(g)$ with $m,n$th entry:

$$\widehat{\Psi}_k(Q_{ml}Q_{nr};g) \equiv \frac{\Phi^*_k(Q_{ml}Q_{nr};g) + \Phi^*_k(Q_{nl}Q_{mr};g)}{\widetilde{\Phi}^*_k(Q_{ml}Q_{nr};g) + \widetilde{\Phi}^*_k(Q_{nl}Q_{mr};g)}. \tag{S17}$$

Finally, for each genetic distance $g$ and each $m,n$ in $T^*$, we average (S17) across all individuals in population $k$. For population pair $m,n$, the vector of (S17) across $g$ (following this averaging) then represents our observed/estimated "coancestry curve" for $m$,$n$ (i.e. this represents our observed value of $\Psi(Q_{ml}Q_{nr};g)$ defined in Equation S9 of Note S3.4). These observed coancestry curves are provided at http://admixturemap.paintmychromosomes.com/ for all analyses presented in this paper. We expect under our theory that the normalisation procedure should mean these coancestry curves decay towards 1 for large $g$ (Note S3).

### S4.3.1 different "grids" of genetic distance bins

We tabulate the counts of $\Phi_k(i, j; g)$, that are used to build our curves, into bins based on three different "grid-types" to capture differing potential ages of admixture, and so scales for ancestry segments:

1. "recent grid" – bins of size 0.1cM for all chunk pairs with $g < 100$cM
2. "ancient grid" – bins of size 0.005cM for all chunk pairs with $g < 5$cM
3. "multiple-date grid" – bins of 0.005cM for all chunk pairs with $g < 5$cM, and bins of size 0.1cM for all chunk pairs with $5 < g < 50$cM.

For all analyses presented here, we initially use the "recent grid." After estimating a single admixture date, the admixture proportion, and the mixing coefficients describing the admixing source groups, using the protocol described in Notes S4.4 and S4.5, we then check to see whether the upper bound of the 95% CI of the estimated date is $> 92$ generations (corresponding to >6.5 half lives of the fitted admixture decay signal, or more exactly to a >99% decay in the admixture signal within 5cM, the range of the ancient grid) or the estimated date is $> 55$ generations (corresponding to >4 half lives). We chose these thresholds to ensure we explore the "ancient grid" when admixture is old enough that this grid will capture almost all the admixture signal. If either is true, we redo the date, proportion and mixing coefficient estimation using the "ancient grid". This "ancient grid" gave a better overall fit than the "recent grid", in terms of coefficient of determination $R^2$, for all simulated populations that met this criterion and for which we observed a signal of admixture. Small deviations from this selection procedure had negligible impact on our results.

In the real data analysis, only three populations showed evidence of admixture that we were both able to characterise and that met the "ancient grid" criterion: Finnish (only when analysed using the "East Europe I" and "East Europe II" analyses described in Note S7), Kalash, and NorthItalian. (The Lahu also showed evidence of ancient admixture, but had an "uncertain" admixture signal – see Note S4.6 – when using either the "ancient" or "recent" grids.) Results for all three were fairly similar when using the "ancient grid" or the "recent grid", e.g. with heavily overlapping date confidence intervals. We report the results for all three populations (and the Lahu) using both the "recent" and "ancient" grids in Tables S12 and S16 but report results using the "recent" grid only elsewhere.

Analogously, when studying "multiple-date" admixture events, if the older date estimate is $> 55$ or has an upper bound of $> 92$ generations, we report results from using the "multiple-date grid". (We perform this additional date estimation step using the final mixing coefficients estimated using the "recent grid" as described in Note S4.5.) Otherwise we report results from using the "recent grid". This "multiple-date grid" gave more accurate date estimates than using the "recent grid" in our simulated datasets with two distinct dates where the second date was 150 generations old (results omitted).

In general, we believe researchers should also visually inspect the observed "coancestry curves" defined in (S17), provided for all of our analyses at
http://admixturemap.paintmychromosomes.com/, in order to ensure appropriate grid choices.

## S4.4 Fitting a single date simultaneously to a group of coancestry curves

The previous section indicates how we construct coancestry curves for a population $k$ based on some current mixture model. We now describe how we use these curves to estimate (initially) a single best-fit admixture date $\lambda$ generations, by fitting an exponential distribution with

unknown rate parameter $\lambda$ to the observed "coancestry curve" vectors generated from (S17). (Unless otherwise noted, we fit to all $g \in [1, 50]$cM when using the "recent grid" or "multiple-date grid" and all $g \in [1, 5]$cM when using the "ancient grid" described in Note S4.3.1.) In Note S4.8, we will consider testing for admixture at multiple times.

According to Equation S11, when admixture occurs at a single time, for all $m, n$ in $T^*$ the coancestry curve is of the form

$$\Psi(Q_{ml}Q_{nr}; g) \equiv \tau_{mn} + \delta_{mn} \exp^{-g\lambda}, \tag{S18}$$

where $\tau_{mn}$ should be equal to 1. We obtain estimators $\hat{\lambda}$ and $\hat{\tau}_{mn}$,$\hat{\delta}_{mn}$ for $m, n \in T^*$ of the respective parameters, to minimize the sum of squared errors:

$$\sum_{m,n \in T^*} \sum_{g} \left( \widehat{\Psi}_k(Q_{ml}Q_{nr}; g) - \tau_{mn} - \delta_{mn} \exp^{-g\lambda} \right)^2. \tag{S19}$$

We use standard least-squares regression and the Nelder-Mead algorithm in the optim function in R. Although we do not restrict it to be so here, we note in practice the inferred $\tau_{mn}$ was nearly always close to 1.0 (see observed coancestry curves at http://admixturemap.paintmychromosomes.com/).

After accounting for units of cM, we interpret $\hat{\lambda}$ as the date in generations from the present that admixture occurred. The predicted (green) lines describing a single admixture event using a pair of donor populations $m, n$ in all figures in the main text, this supplement, and at http://admixturemap.paintmychromosomes.com/ are calculated using $\hat{\tau}_{mn} + \hat{\delta}_{mn} \exp^{-g\hat{\lambda}}$.

To determine confidence bands around the estimated date(s), we generate "pseudo-individuals" for population $k$ by bootstrap resampling of the chromosomes of individuals from population $k$. In particular to generate a single "pseudo-individual", we randomly sample 22 individuals with replacement from $k$, with each individual representing one of chromosomes 1 to 22. We then sum the $\Phi_k(i, j; g)$ (for each $i, j \neq k \in [1, ..., K]$) across the 22 chromosomes of these 22 re-sampled individuals. For each bootstrap re-sample, we generate $n_k$ such "pseudo-individuals" and repeat the steps outlined in (S15)-(S17) and this section to generate a date(s) estimate for that bootstrap re-sample. For all analyses presented here, we used 100 bootstrap re-samples to present confidence intervals.

## S4.5 Iterative procedure to characterise "one-date" admixture and allow testing of admixture hypotheses

In our approach we initially fit a single admixture event between two populations $A$ and $B$, and use the results of this fit to test whether admixture occurred at all, then (if there is such evidence) test whether there is evidence that more complex models involving multiple times of admixture, or multiple admixing groups, are necessary. We describe here how we identify the admixture date and sources for the initial single event fit.

In Note S3.2 we describe how we model haplotypes within admixing populations $A$ and $B$ as mixtures of those found in sampled populations described by mixing coefficients $\vec{\gamma}$ and $\vec{\zeta}$, respectively. Populations $A$ and $B$ mixed $\lambda$ generations ago, contributing proportions of DNA $\alpha$ and $1-\alpha$, respectively, to a current-day population. We also show how given these quantities we could predict overall copying vectors (Notes S3.1-S3.2) and "coancestry curves" for pairs of populations as a function of genetic distance (Note S3.4). We described in Note S4.1.1 how we estimate copying vectors from the data, and use these to fit an initial mixture (Note S4.2) that can be used to produce a set of estimated coancestry curves from the data (Notes S4.3-S4.4).

We attempt to characterize admixture by finding values for $\alpha$, $\lambda$, and the vectors $\vec{\gamma}$, $\vec{\zeta}$ that most closely match the predicted copying vectors and coancestry curves with the observed val-

ues, using a least-squares approach. Because this procedure suggests a new weighting, "cleaned painting" and hence a new set of coancestry curves, we iterate it to attempt to obtain a better fit.

Recall for a simple event, for the admixed population we have an expected copying vector with $j$th element:

$$f_j \;=\; \alpha \sum_{i=1}^{K} \gamma_i f_j^i + (1-\alpha) \sum_{i=1}^{K} \zeta_i f_j^i.$$

Given observed copying vector $\hat{f}^*$, with $j$th element $\hat{f}_j^*$, the sum of squares error (SSE) for a possible admixture representation is then:

$$\sum_{j=1}^{K} \left( \hat{f}_j^* - f_j \right)^2 = \sum_{j=1}^{K} \left( \hat{f}_j^* - \left[ \alpha \sum_{i=1}^{K} \gamma_i f_j^i + (1-\alpha) \sum_{i=1}^{K} \zeta_i f_j^i \right] \right)^2, \tag{S20}$$

which depends on $\alpha$, $\vec{\gamma}$, $\vec{\zeta}$, and where we approximate $f_j^i$ using observed copying vector values (i.e. $\hat{f}_j^{*i}$) for sampled populations $i \in [1, ..., K]$ (we deal with "self-copying" – i.e. copying from members with the same population label – as described in Note S4.1.1).

To capture information from our coancestry curves, we suppose we have some current estimate $\hat{\beta}$ of the vector of overall mixture coefficients $\beta_i = \alpha\gamma_i + (1-\alpha)\zeta_i$ (an initial estimate of this vector is described in Note S4.2), with $n_T$ non-zero coefficients. We also suppose we have used the observed coancestry curves to obtain estimators $\hat{\lambda}$ and $\hat{\tau}_{mn}$,$\hat{\delta}_{mn}$ for $m, n \in T^*$ as described in Note S4.4.

From Equation S11, using the previous definition $D_i \equiv \sqrt{\alpha(1-\alpha)}[Q_i^B - Q_i^A]$ and noting $\beta_i = \alpha\gamma_i + (1-\alpha)\zeta_i$, the underlying coancestry curve for population pair $m$, $n$ has form:

$$\Psi(Q_{ml}Q_{nr}; g) = 1 + \frac{D_m D_n}{Q_m Q_n} \exp^{-g\lambda} = 1 + \delta_{mn} \exp^{-g\lambda}. \tag{S21}$$

where *if the current mixture fit is accurate*, $Q_m = \beta_m$ and similarly $Q_n = \beta_n$. Thus, we formulate an $n_T \times n_T$ matrix $\widehat{H}$ with $m, n$th entry $\widehat{H}_{mn}$:

$$\widehat{H}_{mn} = 2\hat{\beta}_m^k \hat{\beta}_n^k \hat{\delta}_{mn},$$

Note that $\widehat{H}$ is our estimate of the matrix $H$ defined in Note S3.5, and that $\widehat{H}_{mn}$ corresponds to $D_m D_n$ above. (The factor of 2 derives from having to average over phase in diploid individuals, which halves the size of the coefficient.)

To directly estimate $D_i$ for each $i$, we can take an eigen decomposition of the matrix $\widehat{H}$ (after standardizing to make the columns and rows of $\widehat{H}$ sum to 0), extracting the resulting eigenvector with largest eigenvalue. We multiply each entry of the eigenvector by the square root of this eigenvalue; call the resulting $n_T$-vector $\widehat{D}_1$, with the $i$th element denoted as $\widehat{D}_{1,i}$. In general, we analogously let $\widehat{D}_{j,i}$ represent the $i$th element of the eigenvector with $j$th largest eigenvalue, multiplied by the square root of the $j$th largest eigenvalue. Note that $\widehat{D}_{1,i}$ gives our estimate of $D_i$ defined in Note S3.5. Let $\widetilde{H} \equiv \widehat{D}_1 \widehat{D}_1^T$.

To determine the proportion of variance of $\widehat{H}$ captured by $\widehat{D}_1$, we calculate a "fit quality" score:

$$FQ_1 \equiv 1.0 - \frac{\sum_{m=1}^{n_T} \sum_{n=1}^{n_T} (\widetilde{H}_{m,n} - \widehat{H}_{m,n})^2}{\sum_{m=1}^{n_T} \sum_{n=1}^{n_T} (\widehat{H}_{m,n})^2},$$

with $\widetilde{H}_{m,n}$ corresponding to the $m, n$th entry of $\widetilde{H}$. In general, we let $FQ_i$ refer to the "fit quality" measuring the proportion of variance captured by the $i$th eigenvector of the $\widehat{H}$ matrix,

calculated in the analogous way. $FQ_1$ tells us how well a single event fits the coefficient matrix. The additional $FQ_i$ inform us about the additional information captured by independent directions of the eigendecomposition, which can for example be informative about multi-way admixture.

From the definition of $D_j$ we have, if the admixture representation is appropriate:

$$D_j = \sqrt{\alpha(1-\alpha)}\big[\sum_{i=1}^{K}\gamma_i g_j^i - \sum_{i=1}^{K}\zeta_i g_j^i\big],$$

where $g_j^i \equiv \sum_{l=1}^{K} W_{jl} f_l^i$ gives the expected weight for population $j$ using the current weight matrix $W$, for component $i$ in the mixture, so that e.g. $\sum_{i=1}^{K}\gamma_i g_j^i$ is the overall expected value for $Q_{js}$ given the true ancestry is from source $A$ (i.e. $Q_j^A$), as $A$ is described here using mixing coefficients $\vec{\gamma}$.

If the admixture representation is appropriate, $D_j$ and $\widehat{D}_{1,j}$ should be close, and thus

$$\sum_{j=1}^{K}\Big(\widehat{D}_{1,j} - D_j\Big)^2 = \sum_{j=1}^{K}\Big(\widehat{D}_{1,j} - \sqrt{\alpha(1-\alpha)}\big[\sum_{i=1}^{K}\gamma_i g_j^i - \sum_{i=1}^{K}\zeta_i g_j^i\big]\Big)^2 \tag{S22}$$

provides a sum of squares error (SSE) measure of the admixture representation. In practice, we must estimate $g_j^i$ and to do this we use $\hat{g}_j^{*i} \equiv \sum_{l\neq k}^{K}[c_l^k \hat{\beta}_j^k \hat{f}_l^{*j} \hat{f}_l^{*i}]$ with $c_l^k$ defined as in Note S4.3.

To characterize the admixture, we generate initial estimates of $\vec{\beta} \equiv \beta_i \in [1, ..., K]$ as described in Note S4.2, allowing us to produce initial coancestry curves for all pairs of populations included in the mixture. We then seek a mixture representation of the individual sources $A$ and $B$ to minimize (S20) and (S22) simultaneously using a weight $\psi \in [0, 1]$. I.e. we minimize the following:

$$\begin{aligned}&\psi\textstyle\sum_{j\in T^*}\Big(\widehat{D}_{1,j} - \sqrt{\alpha(1-\alpha)}\big[\sum_{i\neq k}^{K}\gamma_i\hat{g}_j^{*i} - \sum_{i\neq k}^{K}\zeta_i\hat{g}_j^{*i}\big]\Big)^2\\ &+(1-\psi)\textstyle\sum_{j=1}^{K}\Big(\hat{f}_j^{*k} - \big[\alpha\sum_{i\neq k}^{K}\gamma_i\hat{f}_j^{*i} + (1-\alpha)\sum_{i\neq k}^{K}\zeta_i\hat{f}_j^{*i}\big]\Big)^2,\end{aligned} \tag{S23}$$

while fixing

$$\psi = \sqrt{\sum_{i\neq k}^{K}(\hat{f}_i^{*k})^2}\Big/\left[\sqrt{\sum_{i\neq k}^{K}(\hat{f}_i^{*k})^2} + \sqrt{\sum_{i=1}^{n_T}(\widehat{D}_{1,i})^2}\right],$$

which weights each of (S20) and (S22) by a term proportional to an approximation of their respective variances.

We aim to infer the proportion of admixture $\alpha$ and the mixing coefficients $\gamma_i$ and $\zeta_i$, $i \neq k \in [1, ..., K]$ that describe the two admixing source groups $A$ and $B$. For each fixed $\alpha$ over a grid of values $[0, 0.01, ..., 0.99, 1.0]$, we perform a non-negative-least-squares (nnls) regression (using the "nnls" package in R) that minimizes (S23) over all $\gamma_i$, $\zeta_i$ under the restrictions that each $\gamma_i > 0$, each $\zeta_i > 0$, $\sum_{i\neq k}^{K}\gamma_i = 1$, and $\sum_{i\neq k}^{K}\zeta_i = 1$. We define $\hat{\alpha}$ as the value that minimizes (S23) over all $\alpha$ using this procedure, and take this $\hat{\alpha}$ to be our estimated proportion of admixture.

Using this fixed $\hat{\alpha}$, we take the nnls-estimated coefficients $\gamma_i$ for $i \neq k \in [1, ...K]$ from (S23) to be our estimates of the mixing coefficients of the first source of admixture (i.e. the source that contributes the proportion $\hat{\alpha}$ of admixture), and the nnls-estimated coefficients $\zeta_i$ for $i \neq k \in [1, ...K]$ from (S23) to be the mixing coefficients of the other source of admixture. We will refer to these estimates as $\hat{\gamma}_i^k$ and $\hat{\zeta}_i^k$, respectively. Taking $\sum_{i\neq k}^{K}\hat{\gamma}_i^k\hat{f}^{*i}$ and $\sum_{i\neq k}^{K}\hat{\zeta}_i^k\hat{f}^{*i}$ gives our model's inferred copying vectors for the two (perhaps extinct) admixing source groups $A$

and $B$. Examples of these inferred source copying vectors are provided in the simulation results of Note S5.

Finally, we re-estimate the overall mixing coefficients describing population $k$, i.e. the $\hat{\beta}$ terms initially calculated as described in Note S4.2, as $\hat{\beta}_i^k = \hat{\alpha}\hat{\gamma}_i^k + (1-\hat{\alpha})\hat{\zeta}_i^k$ for $i \neq k \in [1, ..., K]$. As in Note S4.2, we remove any donor populations $i$ where $\hat{\beta}_i^k \leq 0.001$, defining a new set $T^*$, and re-scale so that the remaining $\hat{\beta}$s sum to 1. We then repeat the generation of coancestry curves and the estimation of the dates as described in Notes S4.3 and S4.4, followed by re-estimating the proportion of admixture and mixing coefficients as described in this section. We repeat this until some convergence criterion is met. In practice, for all results reported here we use 5 total iterations.

For all populations in which we infer simple "one date" events (see Note S4.6 below) reflecting a single admixture event between two source populations, the values of $\hat{\alpha}$ and the $\hat{\gamma}_i^k, \hat{\zeta}_i^k$ after convergence are provided online at http://admixturemap.paintmychromosomes.com/.

For Figure 2 of the main text, we present values of $\hat{\gamma}_i^k$ and $\hat{\zeta}_i^k$ for all depicted populations, after summing across all $i$ in the same "clade" as determined by fineSTRUCTURE (see Table S11 in Note S6.2 for the definition of these clades). These summarise the haplotype sharing of the inferred source groups with the sampled populations. In this figure, we describe the first event (i.e. the most prominent event captured by $\widehat{D}_1$) for populations with "one date, multiway" and the recent event for populations with "two dates" (see Notes S4.6, S4.8 and S4.9). For each admixed population $k$ presented in Figure 2 of the main text, we also show (with colored dots) the "clade" containing the donor population whose copy vector $\hat{f}^{*i}$ is most correlated with $\sum_{j\neq k}^{K} \hat{\gamma}_j^k \hat{f}^{*j}$ among all sampled populations $i \neq k \in [1, ..., K]$, and similarly the "clade" containing the population whose copy vector best matches $\sum_{j\neq k}^{K} \hat{\zeta}_j^k \hat{f}^{*j}$.

## S4.6 Procedure for admixture inference

For each recipient population, we characterize admixture into one of the five following categories, based on results of statistical testing described in the subsequent sections (and after five iterations of the date, proportion, and mixing coefficients estimation algorithm described in Note S4.5):

A. "no admixture" – $p$-value $\geq 0.01$ for the test of admixture described in Note S4.7, implying no strong evidence of admixture within the last $\leq 400$ generations.

B. "uncertain" – $p$-value $< 0.01$, implying evidence of admixture, but $FQ_B < 0.985$, implying either that admixture is complex, involving >3 groups, or our inference is potentially noisy, where $FQ_B$ is defined as follows. As in Note S4.5, let $FQ_j$ refer to the "fit quality" measuring the proportion of variance captured by the eigenvector corresponding to the $j$th largest eigenvalue of an eigen-decomposition of the $\widehat{H}$ matrix defined in Note S4.5. Let $FQ_j^{\text{NULL}}$ denote the corresponding "fit quality" measure for the coancestry curves generated as described in Note S4.7 (and measuring excess ancestry sharing within each individual relative to other individuals from the same population). Let $FQ_B \equiv \min(FQ_1 + FQ_2, FQ_1^{\text{NULL}} + FQ_2^{\text{NULL}})$. This can be interpreted as a measure of the proportion of variance explained by at most three distinct admixing groups. If this measure is not close to 1, i.e. $FQ_B < 0.985$, we do not attempt to infer the admixture history in detail (though we still produce 95% CIs for the admixture time in this setting, which is not expected to be affected). This happened in only a small minority of either our real, or simulated, populations (see Tables S1, S5, S7, S9, S12, S16). Otherwise, we do make an inference, and we assume below that $p$-value $< 0.01$ and $FQ_B \geq 0.985$ suggest our observed signals can be explained by 2 or 3 groups mixing at one or more dates in the past. For populations in which we infer a single date of admixture, we furthermore

check whether the 95% CIs for the date $\lambda$ inferred using our standard approach (Notes S4.4-S4.5) and the approach outlined in Note S4.7 (and measuring excess ancestry sharing within each individual relative to other individuals from the same population) overlap. If they do not, we also classify the admixture as "uncertain", due to simulations and theoretical results suggesting that this can be the sign of a strong bottleneck since the time of admixture affecting inference (see Notes S5.5.3 and S6.4.5 for further details).

C. "multiple dates (2E)" – Using the test described in Note S4.8 we obtain $p$-value $< 0.05$, providing evidence to reject a null hypothesis of admixture at a single time in the past. In this case, we perform inference of admixture at multiple times (we infer two dates) as described in Note S4.8. Otherwise, if the $p$-value $\geq 0.05$, we test for evidence of multiway admixture.

D. "one date, multiway (MW)" – Using the test described in Note S4.9 we obtain $p$-value $< 0.05$, providing evidence to reject a null hypothesis of admixture involving only two groups, at a single time in the past. In this case, we perform inference of 3-way simultaneous admixture, as described in Note S4.9.

E. "one date (1E)" – We regard our data as consistent with "simple" admixture of two groups at a single time in the past if there is evidence of admixture ($p < 0.01$), and neither the multiple-date nor multiway tests provide evidence to reject such a scenario ($p \geq 0.05$). Inference is performed as described in Note S4.5.

In addition to these automated tests, we believe it is also highly worthwhile to examine the coancestry curves (e.g. to verify signal strength, and visualise evidence of complex admixture and verify this is consistent with the conclusions reached), as in Figures 2-4 of the main text. All such curves for our real data analysis are provided at
http://admixturemap.paintmychromosomes.com/.

## S4.7 Determining $p$-values for evidence of admixture

The procedure above generates coancestry curves for a set of donor groups inferred as important for representing the unknown admixing sources. These coancestry curves provide evidence of admixture in the past, if they show a decay with genetic distance, indicating local correlation in genetic ancestry, over and above that expected by chance. Estimated admixture dates of 1 generation (the minimum value), or a very large number of generations (we use $\geq 400$ generations, so that correlation will disappear at very small genetic distances, smaller than those we use for the coancestry curves) imply no decay with genetic distance, and so no evidence of admixture. Thus, we test for evidence of admixture by asking the proportion of curves in 100 bootstrapped samples that show an estimated admixture date of 1, or $\geq 400$, generations, to obtain a bootstrapped $p$-value of no admixture.

To construct a robust test, we use the idea that in truly admixed populations, ancestry segments producing "admixture LD" occur within individual genomes, resulting in ancestry LD characteristically decaying within individual genomes, much more strongly than when ancestry is measured in different individuals. To construct a test of this based on our method's inference of genetic ancestry, we first generate "across-individual" coancestry curves, by considering CHROMOPAINTER painting samples from *different* individuals, and use these to normalise our original coancestry curves. Specifically, to assess admixture in group $k$ that has $n_k$ sampled individuals, using the notation of Note S4.3, we take $\vec{X}^a$ from one individual sampled from $k$ and $\vec{X}^b$ from a different individual sampled from $k$. For results presented here, we select one painting sample to represent each haplotype of each individual in $[1, ..., n_k]$. We then calculate $\Phi_k(i, j; g)$ in Equation S14 by comparing each of these individuals' haplotypes to the two painting samples

of each other individual in $k$, using these new definitions of $a$ and $b$. Following precisely the remaining equations in Note S4.3, we calculate a new value of $\widehat{\Psi}_k(Q_{ml}Q_{nr};g)$ (from Equation S17) using this new $\Phi_k(i,j;g)$; call this new value $\widehat{\Psi}_k^{\text{NULL}}(Q_{ml}Q_{nr};g)$. Finally, for each $m,n \in T^*$ and $g$, we take our observed "coancestry curve", defined in Note S4.3 to be the average of $\widehat{\Psi}_k(Q_{ml}Q_{nr};g)$ across all individuals in $k$, and divide it by $\widehat{\Psi}_k^{\text{NULL}}(Q_{ml}Q_{nr};g)$.

We next fit this new curve, i.e. defined by $\widehat{\Psi}_k(Q_{ml}Q_{nr};g)/\widehat{\Psi}_k^{\text{NULL}}(Q_{ml}Q_{nr};g)$, exactly as described in Notes S4.4 and S4.5, using five iterations of alternating date and proportion estimation, assuming a single admixture event, and using the "recent grid" defined in Note S4.3.1. Following the five iterations, we perform 100 bootstrap re-samples of individuals' chromosomes to infer new values of $\widehat{\Psi}_k(Q_{ml}Q_{nr};g)$. We then divide each of these in turn by $\widehat{\Psi}_k^{\text{NULL}}(Q_{ml}Q_{nr};g)$ and re-infer dates as described in Note S4.4.

To assess evidence of *any* admixture, we obtain an empirical $p$-value as $p = D/101$, where $D$ is the number of bootstraps (including the original data) with a date $\lambda$ where $\lambda \leq 1$ or $\lambda \geq 400$, rejecting the null of no admixture only if $p < 0.01$.

## S4.8 Multiple dates of admixture

In populations with evidence of admixture according to the previous two sections, we perform a test as to whether there is evidence that admixture occurred at more than one time, by asking whether a model with a mixture of decay rates (we stop at two rates, though these ideas could be extended to additional dates) better fits our coancestry curves than a nested null model allowing a single decay rate. We utilise properties of the coancestry curves under multi-way admixture derived in Note S3.6.

Let $R_1^{(m,n)}$ refer to the coefficient of determination (i.e. $R^2$) value for population pair $m,n \in T^*$ from the fit of (S19) in Note S4.4, assuming a single date of admixture.

Similarly let $R_2^{(m,n)}$ refer to the coefficient of determination value for a 2-date model allowing two decay rates, which we fit as follows. To estimate the two dates of admixture, we replace (S19) in Note S4.4 with the following, which predicts the observed coancestry curves defined by $\widehat{\Psi}(Q_{ml}Q_{nr};g)$ with the sum of two exponential distributions with different rates $\lambda_1$ and $\lambda_2$:

$$\sum_{m,n \in T^*} \sum_{g} \left( \widehat{\Psi}(Q_{ml}Q_{nr};g) - \tau^*_{mn} - \delta_{1;mn} \exp^{-g\lambda_1} - \delta_{2;mn} \exp^{-g\lambda_2} \right)^2. \tag{S24}$$

Analogous to the single date case, we minimize (S24) over $\lambda_1, \lambda_2$ and $\tau^*_{mn}, \delta_{1;mn}, \delta_{2;mn}$ for $m,n \in T^*$, denoting their respective estimated values $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\tau}^*_{mn}, \hat{\delta}_{1;mn}, \hat{\delta}_{2;mn}$. Note that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ represent our estimates of each event's date. To avoid "non-sensical fits" (often with two almost identical dates that nearly cancel), we disallow rate parameters where for some $m$, $\delta_{1;mm}$ and $\delta_{2;mm}$ are of opposite sign (which cannot happen in theory) beyond the amount likely due to fitting noise, so that one of $\delta_{1;mm}, \delta_{2;mm}$ is greater than 3 times the standard deviation of the fitted model residuals for the right-most 50% of the vector of values across $g$. The predicted (green) lines describing two admixture events with different dates using a pair of donor populations $m,n$ in all figures in the main text, this supplement, and at http://admixturemap.paintmychromosomes.com/ are calculated using $\hat{\tau}^*_{mn} + \hat{\delta}_{1;mn} \exp^{-g\hat{\lambda}_1} + \hat{\delta}_{2;mn} \exp^{-g\hat{\lambda}_2}$. See Figure S2 for an example of one of our sampled populations (Sindhi) fit with one versus two dates.

Given the fitted $\hat{\lambda}_1$ and $\hat{\lambda}_2$ we obtain $R_2^{(m,n)}$, the usual coefficient of determination value for the 2-date model, for each pair of curves. For those pairs of populations $T^*$ whose curve is not already almost perfectly fit by a single date ($R_1^{(m,n)} < 0.975$), we calculate $M = \max_{m,n \in T^*}(\frac{R_2^{(m,n)} - R_1^{(m,n)}}{1.0 - R_1^{(m,n)}})$. This measures the maximum fraction of additional variance explained

by adding a second date, across the curves we produce. For a single pair of curves $(m, n)$ and a fixed number of points used to construct the coancestry curve, this measure allows calculation of the ratio of residual sum of squares allowing one or two curve rates respectively, and so is equivalent to the test-statistic for a standard likelihood ratio test of multiple decay rates (with an approximate asymptotic chi-squared null distribution), assuming i.i.d. normal errors in the coancestry curves. Increasing values of $R_2^{(m,n)}$ indicate stronger evidence against the null. This makes this measure $M$ natural, and maximising across $m, n \in T^*$ allows detection of evidence of multiple dates from any pair of inferred groups in the donor set.

However, the use of multiple groups and lack of independence of points in the coancestry curves mean we cannot directly use the standard chi-squared null distribution. Instead, we assign an empirical $p$-value to an observed value of $M$ using 4,747 one event "real-sample" simulations (therefore simulated under a null model of a single date; we note this is conservative to the potential presence of additional date signals due to our use of real samples) described in Notes S5.1-S5.2. For the real analysis and "coalescent-based" simulations described in Notes S5.4-S5.5, we report $p$-values for concluding "multiple dates" by tabulating the proportion of these simulations whose values of $M$ are as or more extreme than that of the given group, using a 5% threshold for rejection of the null hypothesis of a single date. For populations showing evidence of multiple dates according to this test, we confirm that the re-normalised curves described in Note S4.7 also show evidence of multiple dates, requiring that $M$ calculated from these curves is reduced by no more than 2/3, and otherwise accept the null. (This robustness check only influenced the MbutiPygmy in our real population analyses, though it did influence some of the "coalescent-based" simulations described in Notes S5.4-S5.5 – we highlight such populations by removing their multiple date $p$-value.) We note that this procedure (as any) might lack power to detect multiple admixture dates in some settings, and explore power using simulations in Notes S5.1-S5.2 and Notes S5.4-S5.5.

For populations whose conclusion is "two dates", if the point estimate of the older event date is $> 55$ generations or has an upper bound $> 92$ generations, we re-estimate both dates (as well as CIs based on bootstrap re-sampling) using the "multiple-date" grid, keeping the mixing coefficients inferred using the "recent grid" (see Note S4.3.1 for details of the various grid types).

Analogous to the definition of $\widehat{H}$ in Note S4.5, in the two date scenario we consider $\widehat{H}^1$ and $\widehat{H}^2$ to capture information on the recent and older event, respectively, with the $m, n$th components of each equal to:

$$\widehat{H}^1_{mn} = 2\hat{\beta}^k_m \hat{\beta}^k_n \hat{\delta}_{1;mn}$$
$$\widehat{H}^2_{mn} = 2\hat{\beta}^k_m \hat{\beta}^k_n \hat{\delta}_{2;mn}$$

Starting with characterizing the most recent event, we substitute $\widehat{D}_1$ in (S23) with the analogous value from an eigen-decomposition of $\widehat{H}^1$. We then infer $\alpha$, $\gamma_i$, and $\zeta_i$ for $i \neq k \in [1, ..., K]$ in the same manner as described in Note S4.5.

However, in contrast to the simple admixture case, we next calculate:

$$\varrho_j \equiv \sum_{i \neq k}^{K} c^k_\varrho (\hat{\gamma}^k_i - \hat{\zeta}^k_i) \hat{f}^{*i}_j / (n_j - 1) \tag{S25}$$

for all $j \in [1, ..., K]$, with $c^k_\varrho \equiv 1.0 / \sum_{j=1}^{K} [\sum_{i \neq k}^{K} (\hat{\gamma}^k_i - \hat{\zeta}^k_i) \hat{f}^{*i}_j])$. The $\varrho_j$ give our inferred *differences* in the copying vectors of the two admixing sources (i.e. analogous to $D_i / \sqrt{\alpha(1-\alpha)}$ as defined in Note S3.5), which are expected to be more stably estimated than the mixing coefficients $\gamma_i, \zeta_i$ (see Note S3.5). Dividing each $\hat{f}^{*i}_j$ by $n_j - 1$ keeps populations with more sampled individuals from having larger $\varrho_j$.

For *all* populations depicted in Figures S16-S21 in Note S7, no matter what their conclusion (A)-(D) from Note S4.6, we plot $|\varrho_j|/[\sum_{i=1}^{K}|\varrho_i|]$. However, for these figures we replace the $\hat{f}^{*i}$ in Equation S25 by a new set of vectors that sum $\hat{f}_j^{*i}$ across all $j$ in the same "clade" as determined by fineSTRUCTURE (see Table S11 in Note S6.2 for the definition of these clades).

Otherwise for populations with "multiple dates", we take the ten most positive $\varrho_j$ from (S25) and the absolute value of the ten most negative $\varrho_j$ to illustrate the mixing coefficients for the two admixing source groups, respectively. In particular we rescale each set of ten values to sum to 1 and multiply them by $\hat{\alpha}$ and $1-\hat{\alpha}$, respectively. Call these rescaled values $\hat{\varrho}_j$ for $i \in [1, ..., K]$, with the properties that at most 20 $\hat{\varrho}_j$ are greater than 0, all other $\hat{\varrho}_j$ are equal to 0, and $\sum_{j=1}^{K}\hat{\varrho}_j = 1$. We furthermore set $\hat{\beta}_j^k = \hat{\varrho}_j$ and use the procedure described in Note S4.3 to generate observed coancestry curves for all pairwise combinations of donor populations with $\hat{\varrho}_j > 0$. We provide $\hat{\varrho}_j$ and these coancestry curves for all "multiple date" populations at http://admixturemap.paintmychromosomes.com/.

We repeat this procedure when characterizing the older admixture event, instead taking an eigen-decomposition of $\widehat{H}^2$.

## S4.9 Multiple simultaneous admixture events

In populations with evidence of admixture, but no strong evidence ($p > 0.05$) this occurred at more than one time, we perform a test of whether admixture involved more than two groups, using a similar approach to the previous section, but based on the $\widehat{H}$ matrix defined in Note S4.5, relating to the inferred intercepts of our coancestry curves (i.e. values of these curves at genetic distance zero). As shown in Note S3.6, if $G$ groups simultaneously admix, this matrix is predicted in theory to have rank $G-1$, and so testing the number of admixing groups reduces to a problem of testing the rank of the underlying matrix $H$. In a model where errors in $\widehat{H}$ are independent normal (we will relax this assumption), we note that $FQ_1$ defined in Note S4.5 gives the proportion of variance explained by the least squares (i.e. maximum likelihood) rank-1 estimator of $H$. Similarly $FQ_2$ gives the additional proportion of variance explained by adding a second rank (or 3 admixing groups), i.e. $FQ_1 + FQ_2$ gives the variance explained by the least squares rank-2 estimator of $H$, and so on.

We do not feel able to analyse groups where obviously $G \geq 4$, or where $\widehat{H}$ may be noisily estimated, and so characterise populations as "uncertain" whenever the rank-2 estimator does not almost completely explain (>98.5%) the variance of $H$. For each $j$, let $FQ_j^{\text{NULL}}$ denote the analogous "fit quality" measure for the normalised coancestry curves generated as described in Note S4.7. As stated in Note S4.6, letting $FQ_B \equiv \min(FQ_1 + FQ_2, FQ_1^{\text{NULL}} + FQ_2^{\text{NULL}})$, for robustness we call a population as "uncertain" if $FQ_B \leq 0.985$. Otherwise, $\widehat{H}$ is very well approximated by a matrix of rank at most 2, corresponding to $\leq 3$ admixing groups.

We chose to use as a natural measure of the strength of evidence for multiple admixing groups (although other natural measures could also be used) the fraction of variance explained by adding a second group, i.e. $FQ_2 \approx 1 - FQ_1$. We thus obtained a $p$-value for testing a null of a a simple two source admixture event by the fraction of values of $FQ_1$ in our 4,747 one event "real-sample" simulations where admixture was concluded that are below that observed, analogous to the empirical $p$-value approach of the previous section. (This use of simulations allows us to account for the departures from normality and independence in our real data.) As in the previous section, we reject the null if this $p < 0.05$.

In populations where this null is rejected, a single event does not adequately describe the data, because we see distinct "directions of variation" based on our curve fits (i.e. using the metrics defined in Note S4.6). For such populations, we have already inferred the date of admixture from the procedure described in Note S4.4. To infer proportions of admixture and mixing coefficients, we follow the same procedure described in Note S4.5 to describe one of the events, thus

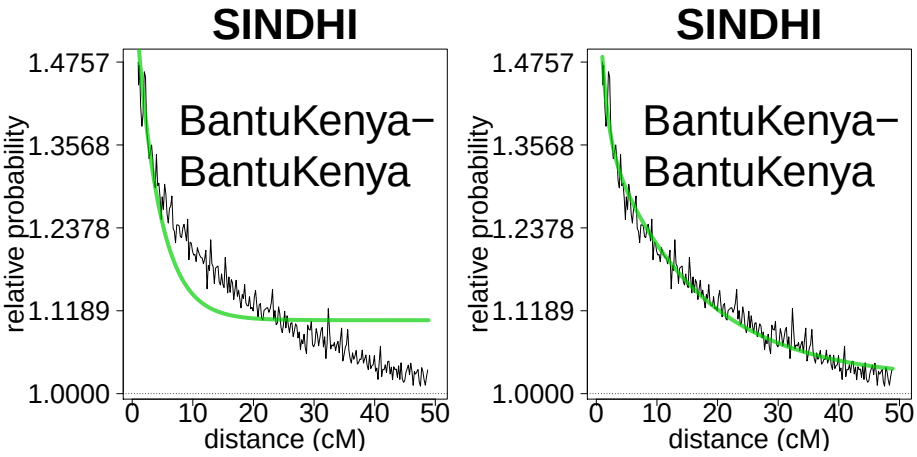



**Figure S2:** Fitting a single exponential distribution (left) versus the sum of two independent exponential distributions with different rates (right) to coancestry curves in the Sindhi population (under the "Central Asia" analysis described in Note S7.4). Black lines represent the (re-scaled) probabilities (i.e. from Equation S17) of copying from one Bantu Kenya segment to another as a function of increasing genetic distance (in cM) between the two segments. Green lines illustrate the fit of the exponential distribution(s) with rate(s) inferred as described in Notes S4.4 and S4.8.

corresponding to the leading eigenvalue/eigenvector. To describe the second event, we replace $\widehat{D}_1$ with $\widehat{D}_2$, the corresponding vector for the second largest eigenvalue/eigenvector, in (S23) but otherwise repeat the same procedure. As in the multiple date scenario, for each event we generate the final observed coancestry curves and differences in copying vectors $\hat{\varrho}_j$ reported at http://admixturemap.paintmychromosomes.com/ using the same procedure described in Note S4.8.

As in the multiple-date case, it is difficult to characterize these multi-way events fully from the coancestry curves and copying vectors. Any two vectors spanning the same subspace as the two principal eigenvectors could be viewed as a possible admixture "direction", and so there is uncertainty in the true nature of events. Thus our representation is analogous to a "basis" for admixture, chosen so that the first inferred direction explains (in a sense) as much of the overall admixture signal as possible. Similarly, the second inferred direction explains as much of the remaining signal as possible. However, other possible directions with different splits of the signal are likely to be equally plausible, while we use differences in inferred copying vectors to capture the components (and nature) of these different directions. Approaches that directly infer chunks - and increases in available data - might help such cases in the future.

# S5 Simulations

## S5.1 Details of "real-sample" simulations

To test our method's performance, we performed several sets of simulations. For each simulated population defined below, we generated $n$ independent simulated individuals using phased data from haploids of two different real data populations $A$ and $B$ from our dataset (for phasing details, see Note S6.1). Each simulated admixed individual was generated as a mosaic of segments drawn from our real dataset. Diploid individuals were constructed by aggregating two haploids.

Each simulated admixed haploid of an individual was generated in the following manner, which closely follows previous work (48; 4; 11). First a centimorgan genetic distance $x$ was sampled from an exponential distribution with rate $\lambda/100$, with $\lambda$ corresponding to the time in generations since the admixture event. Then the first $x$ cM of the simulated haploid was composed of the first $x$ cM of a real data haploid from population $A$ with probability $\alpha$, otherwise it was composed of the first $x$ cM of a real data haploid from population $B$. A new genetic distance was sampled from the same exponential($\lambda$) distribution, and this process was repeated until the entire simulated haploid was generated.

To limit the chance of multiple simulated individuals copying from the same real data individual at any area of the genome, we mimicked (4) in that wherever possible the new haploid sampled was selected from the pool of haploids in the selected population $A$ or $B$ for which no other previously simulated haploid had copied at the same location. When this was not possible, a haploid was selected at random from the selected population $A$ or $B$. (We note that an alternative simulation strategy where we sampled randomly from the haploid pool of $A$ or $B$ at each breakpoint did not noticeably affect parameter estimation, e.g. the estimate of the date $\lambda$.)

### S5.1.1 "one-date simulations"

For the "one-date simulations", we simulated admixture between the following pairings of populations $A$ and $B$:

1. Brahui, Yoruba ($n = 20$ samples)
2. Brahui, Han ($n = 20$)
3. French, Brahui ($n = 20$)
4. Colombian, Han ($n = 7$)
5. Yoruba, French ($n = 20$)

For each population pairing 1-5, we simulated admixture in three different proportions ($\alpha$) from the second listed source: 5%, 20% and 50%, and at three different times ($\lambda$): 7, 30 and 150 generations in the past. This results in $5 \times 3 \times 3 = 45$ simulations in total for all possible combinations. We designed these initial simulations to represent a cross-section of scenarios ranging from easier to more difficult, with lower admixture fractions and perhaps more ancient admixture likely to be more difficult to detect. In addition, admixture between more diverged populations (e.g. Yoruba and French) is likely to be easier to detect than comparable admixture between less diverged groups (e.g. French and Brahui). The fractions, dates and populations involved in these simulations also (we believe) do a reasonable job of covering events detected using our approach in the actual data.

#### S5.1.2 "no-admixture simulations"

We furthermore simulated individuals with "no-admixture" by sampling haploids from the same or genetically similar populations admixed at a very distant time in the past (i.e. $\lambda = 1000$). This simulated admixture is so ancient (i.e. such that admixture segments are expected to be tiny), and between such similar groups, that we believe our methods should not detect any admixture signal. We refer to these 5 simulated populations as the "no-admixture simulations":

6. French, French, $\alpha = 0.50$, $\lambda = 1000$ ($n = 20$)
7. Brahui, Brahui, $\alpha = 0.50$, $\lambda = 1000$ ($n = 20$)
8. Han, Japanese, $\alpha = 0.50$, $\lambda = 1000$ ($n = 20$)
9. Sindhi, Pathan, $\alpha = 0.50$, $\lambda = 1000$ ($n = 20$)
10. Basque, French, $\alpha = 0.50$, $\lambda = 1000$ ($n = 20$)

#### S5.1.3 "half-admixture simulations"

We also performed the following simulations where only 50% of the individuals in the simulated population were admixed and the rest had "no-admixture", mainly in order to test our robustness to the scenario of a heterogenous population. We note that in our real analysis, we attempt to mitigate this issue via the use of fineSTRUCTURE to cluster individuals (initially without using labels) into genetically homogenous groups, followed by removal of outliers (see Note S6.2).

11. Brahui, Yoruba "half-admixture":
    - 10 individuals taken from simulation scenario 7
    - 10 individuals simulated with 80% Brahui and 20% Yoruba admixture (taken from previous "one-date simulations")
12. French, Brahui "half-admixture":
    - 10 individuals taken from simulation scenario 6
    - 10 individuals simulated with 80% French and 20% Brahui admixture (taken from previous "one-date simulations")

For each of simulation scenarios 11 and 12, three different simulated populations were made according to whether the admixture between the Brahui and Yoruba (respectively French and Brahui) occurred $\lambda = 7$, 30, or 150 generations ago. This gives 6 simulated "half-admixture" populations in total.

#### S5.1.4 "two-date simulations"

Finally, to test our ability to understand more complex admixture, we simulated four "two-date" populations of $n = 20$ individuals each that have experienced two separate admixture events occuring at different times. For these four simulations, we simulated a recent admixture event occurring $\lambda_1 = 7$ generations ago with an $\alpha_1 = 20\%$ contribution from the Yoruba, and the remaining admixture from one of the four simulated populations listed in 13-16 below. Each of these latter populations has already experienced an older single admixture event occurring at time $\lambda_2$ between the two source populations listed below, with $\alpha_2$ of the admixture contributed from the second listed population:

13. Brahui, Han, $\alpha_2 = 0.2$, $\lambda_2 = 37$
14. Brahui, Han, $\alpha_2 = 0.5$, $\lambda_2 = 37$
15. Brahui, Han, $\alpha_2 = 0.2$, $\lambda_2 = 157$
16. Brahui, Han, $\alpha_2 = 0.5$, $\lambda_2 = 157$

#### S5.1.5 data analysis for "real-sample" simulations

In total we generated 60 simulated populations over the 16 different scenarios described above. In order to assess our power of identifying and describing admixture events, for each of the 60 simulations we performed 100 bootstrap re-samplings of individuals' chromosomes to provide a further 100 approximately independent simulations. In each bootstrap, we re-sampled chromosomes until we had generated $n$ "new" individuals, and performed the identical analysis procedure to that described below. Along with the original (i.e. pre-bootstrapped) simulation, this gave us $60 \times 101 = 6,060$ total simulated datasets evaluated in this section.

When analysing each simulated population, the source populations used to simulate the data were excluded from the potential set of donor populations. This mimics our real data analysis, where the sampled donor populations are at best imperfect matches of the original admixing source groups, for example due to sampling scheme or the original sources being extinct. In this manner we can evaluate our model's ability to reconstruct the genetic make-up of these source populations, e.g. by comparing the copying vectors that our model predicts for each source to those of the actual source populations used to generate them. We note that the populations used in our simulations vary in how closely they are "approximated" by another single sampled group after their removal, with for example the Balochi being genetically closely related to the Brahui (fineSTRUCTURE was unable to separate these groups well), while the Colombians and Yorubans do not have such close proxies. This may also affect our power to identify admixture, and is certainly expected to influence the number of populations required in our mixture representation of the sources (e.g. Figure 1D of the main text).

To more accurately mimic our "full analysis" (see Note S6.3), for each simulated population we would have to re-paint every donor individual in a manner that allows them to copy from the simulated population and also precludes them from copying from the admixing source populations used to generate the simulated data. To avoid this major computational burden, we instead fixed the amount that each donor population copies from the simulated dataset at 0. Furthermore, we removed from each donor population's copying vector (inferred as described in Note S4.1.1) the proportion of admixture copied from each of the true admixing source populations, and rescaled these vectors to sum to 1. These two shortcuts may result in a slight loss of power compared to the real data "full analysis" protocol, for example because the latter shortcut implicitly assumes that every donor population would copy from these admixing source populations equally.

To mimic the fact donor groups were not allowed to copy from the simulated population, we also set the amount each simulated individual copies from the other simulated individuals to 0 (i.e. we disallowed any "self-copying"). Furthermore, to generate copying vectors and painting samples for the simulated groups, we allowed each simulated individual to copy from all individuals from each donor population rather than leaving one individual out of each donor group (i.e. as was done to infer donor copying vectors – see Note S4.1.1). We have found that excluding one individual from each donor group makes little difference relative to including all individuals in our applications here, in that often the relative overall amount copied from the entire donor group will be very similar (results omitted). When generating these painting samples and copying vectors for simulated individuals, we also used estimates of $N_e$ and $\theta$ from the "full analysis" of the real data weighted by the true proportions of simulated admixture,

rather than estimating these values directly. For example, for simulations with 80% Yoruba and 20% French, we took our $N_e$ (respectively $\theta$) estimate to be 0.8 times the $N_e$ (respectively $\theta$) estimate averaged across Yoruba individuals in the "full analysis" of the real data, described in Note S4.1.2, plus 0.2 times the $N_e$ (respectively $\theta$) estimate averaged across French individuals in the "full analysis". To avoid a futher computational burden, for each simulation we used the same copying vector, inferred by running CHROMOPAINTER using the original simulation as a recipient, to represent that simulation across all 100 of its bootstrap replicates.

Outside of these minor discrepancies, we used the identical protocol when analyzing the real data, as outlined in Notes S4.1-S4.6.

## S5.2 Results of "real-sample" simulations – assessing power

For each of the 60 simulations, we first assessed whether the original (i.e. non-bootstrapped) simulation showed any evidence of admixture, i.e. by calculating its empirical $p$-value as described in Note S4.7. Encouragingly all five of our "no-admixture" simulations (Note S5.1.2) did not falsely detect admixture, all with empirical $p$-values $> 0.25$. Only four of our remaining 55 simulations ($\approx$7%) were unable to detect real admixture ($p$-values ranging from 0.05 to 0.36), these being very difficult cases of admixture events occuring 150 generations ago with a proportion of 20% or less between the French and Brahui or the Colombian and Han. For these nine simulations concluding "no admixture" we present no further results beyond their empirical $p$-value in Table S1.

We next considered the 51 of our populations simulated with a single event, i.e. those described in Notes S5.1.1 and S5.1.3, in order to identify thresholds to control our (empirical) type I error rates for falsely calling multiple-dates or multiway admixture. Excluding the four simulations that falsely concluded "no admixture" above, 47 of these 51 showed evidence for admixture ($p$-value $< 0.01$). Including the 100 bootstrap re-samples for each of these 47 groups, this gave us 4,747 total simulated datasets. Each were analysed as described in Note S4 using the "recent" grid (see Note S4.3.1). These analyses provided 4,747 values of $M \equiv \max_{m,n \in T^*}(\frac{R_2^{(m,n)} - R_1^{(m,n)}}{1.0 - R_1^{(m,n)}})$ and $FQ_1$, the parameters we use to assess evidence of "multiple dates" and "one date, multiway" admixture, respectively (see Notes S4.8-S4.9). We used these 4,747 values to obtain empirical $p$-values for our tests of admixture at more than one time, and of multi-way admixture, as described in Note S4.6, and thus thresholds for rejection at the 5% level for each test.

These thresholds guarantee a 5% rejection rate overall for each test. However, we can nonetheless assess the relative performance of these tests for each of the 51 simulation scenarios that detected any presence of admixture. Thus as with our real data analysis, for each of these 51 simulation scenarios we characterized the admixture for each of its 101 replicates into one of the four categories B-E described in Note S4.6. We provide the proportions of these 101 replicates whose conclusion was "uncertain", "one date", "multiple dates", and "one date, multiway" in Table S1 and Figure S7. Furthermore, for all replicates that did not conclude "uncertain" admixture, we report the inner 95% range for both GLOBETROTTER's inferred proportion of admixture and the $F_{ST}$ (49) between the "best match" (assessed by correlation coefficient $r$) to GLOBETROTTER's inferred copying vector for each source and its corresponding true admixing source, in order to assess our uncertainty regarding this proportion and the populations involved in admixture. In the case of the "two event simulations" of Note S5.1.4, we provide these values only for replicates that concluded "multiple dates" when describing results for the second event. The point estimate for the dates in Table S1 and Figure S7 are for the original (i.e. non-bootstrapped) simulation and the CI is based on 100 bootstrap re-samples. After following the protocol outlined in Note S4.3.1, all single event simulations with $\lambda = 150$ and where admixture was detected used the "ancient grid" for infering these

dates. Example results from our method's application to four particular simulated populations are summarized in Figures S3-S6.

For the 47 "one event simulations" that infer any admixture, 26 (55%) correctly infer one-event across all 101 replicates, with an additional 11 inferring one-event in over 90% of their replicates and 3 more in over 75% of their replicates. Among these 40 cases (85% of the 47), the date, proportion and source estimation is extremely accurate when describing the admixture (see Table S1 and Figure S7). There are 6 remaining simulations that falsely infer "multiple dates" or "one date, multiway" admixture in over 25% of their bootstrap replicates, all of which are cases with only 5-10% admixture (this includes 5 of the 13 cases with only 5% admixture where our model finds any evidence for admixture, i.e. ≈40% of such cases). Furthermore, all three cases with "multiple dates" false inference occuring over 25% of the time involve instances simulated with ≈20-95% Brahui and involve coancestry curves corresponding to African donor populations, and thus we believe it is plausible that these false positives may actually reflect GLOBETROTTER picking up real African admixture in the Brahui population. (We note that we do observe a relatively small amount of African admixture in our "real-sample" analysis of the Brahui. Thus in the sense that real populations might have additional true admixture signals, our approach is conservative by penalising conclusions inferring admixture events that differ in any way from those we directly simulate.) Finally, the admixture is "uncertain" >20% of the time for only one of the 47 simulation scenarios, which similarly is a difficult case of 5% admixture contributed 150 generations ago.

Two of the four "two event simulations" conclude "multiple dates" across all replicates, with the other two concluding "multiple dates" 99% and 89% of the time. Source estimation is always perfect and proportion estimation extremely reliable for the recent event. This is most often the case when describing the older event as well, though GLOBETROTTER has difficulty always capturing the minor source perfectly in the old event in our most difficult case. For this reason, we use a slightly different, more robust protocol to describe source groups for the older or least strong events in "multiple date" and "one date, multiway" admixture cases as described in Notes S4.8 and S4.9.

We note that across all 51 simulations, our date esimation is extremely reliable regardless of source and proportion inference, with the true simulated date lying (just) outside the 95% CI in only two of 55 inferred events (3.6%). That suggests our date estimation should be reliable even if admixture proportions or sources have been misclassified.

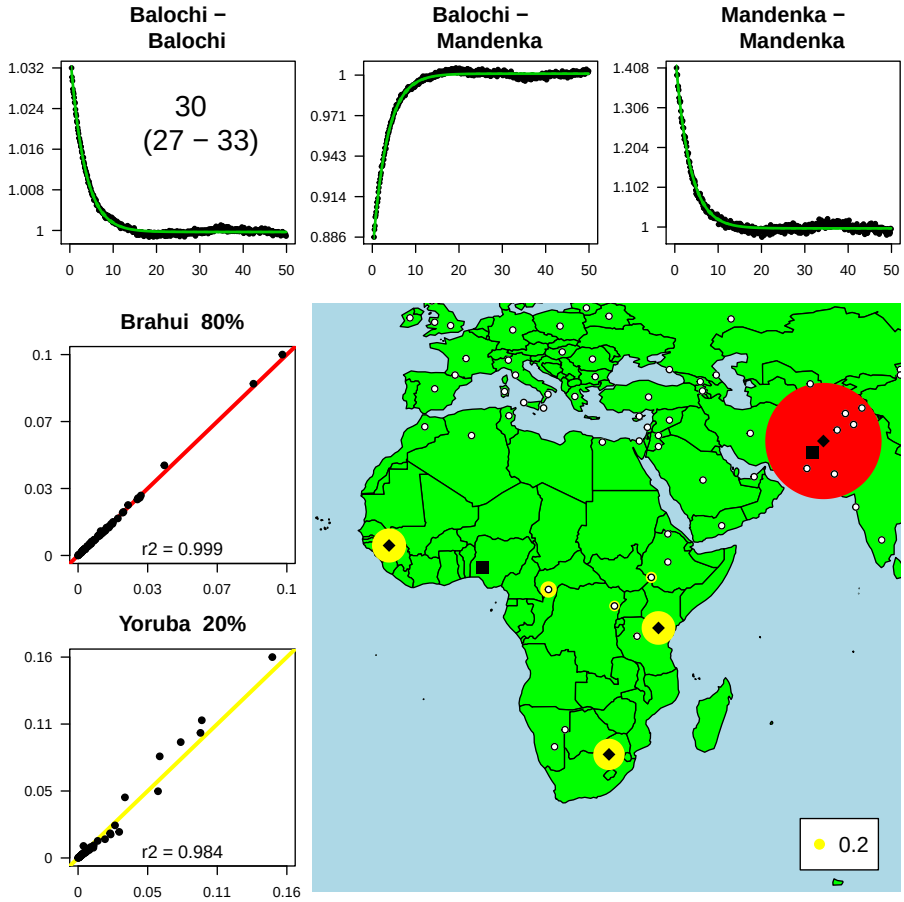



**Figure S3:** Results for simulations with 20 individuals admixed at 30 generations ago, with 80% of their DNA from Brahui and 20% from Yoruba. The top row gives the observed coancestry curves (black lines) for donor populations (Balochi, Mandenka) inferred to represent source groups on opposite sides of event, with fitted exponential distributions based on our inferred date in green. Our inferred date (+ 95% CI) is given in the top left plot. Underneath this at left are the individual-averaged copy vectors, as inferred by our model, versus the true copy vectors for the Brahui (middle left) and Yoruba (bottom left) source populations, with line of equality across the diagonal. In the map at bottom right, the squares give the locations of the true admixing source populations, with white circles denoting all other (potential donor) populations within the region. Diamonds indicate populations among these that have $F_{ST}$ within 0.001 of the minimum $F_{ST}$ across all populations to either of the true admixing sources. Circles depict donor populations inferred to represent each source, with sizes proportional to the inferred mixing coefficients. The total estimated proportion from the yellow source is given in the legend.

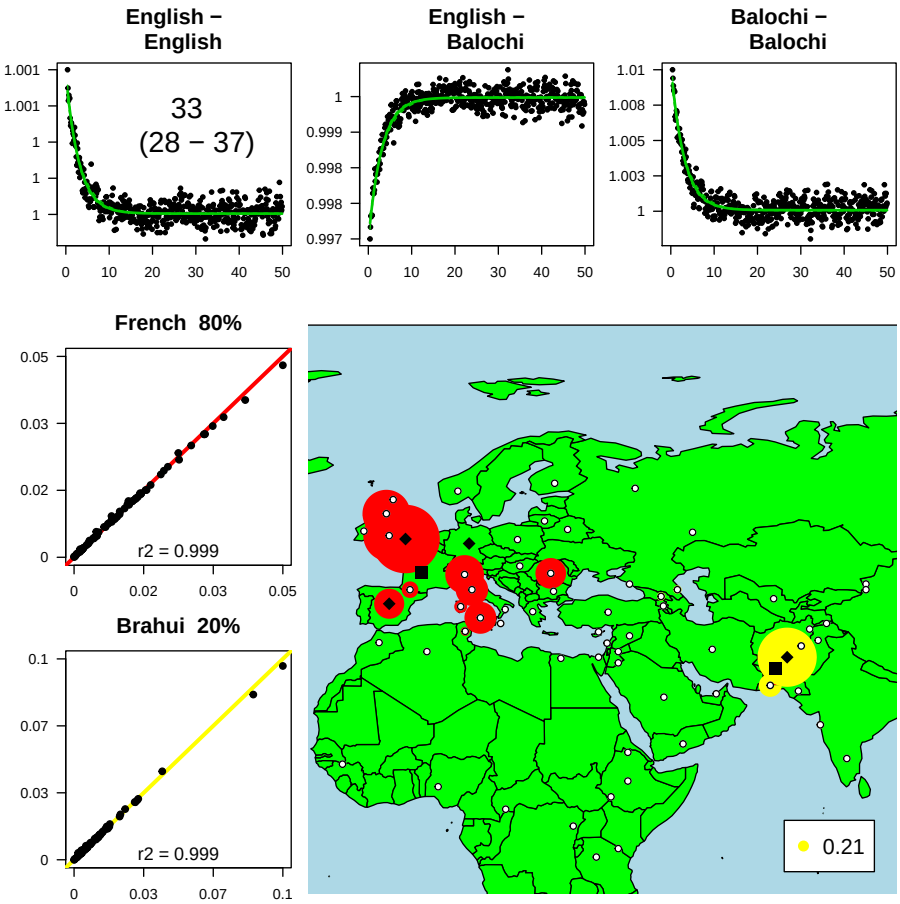



**Figure S4:** Results for simulations with 20 individuals admixed at 30 generations ago, with 80% of their DNA from French and 20% from Brahui. Plot details as in Figure S3.

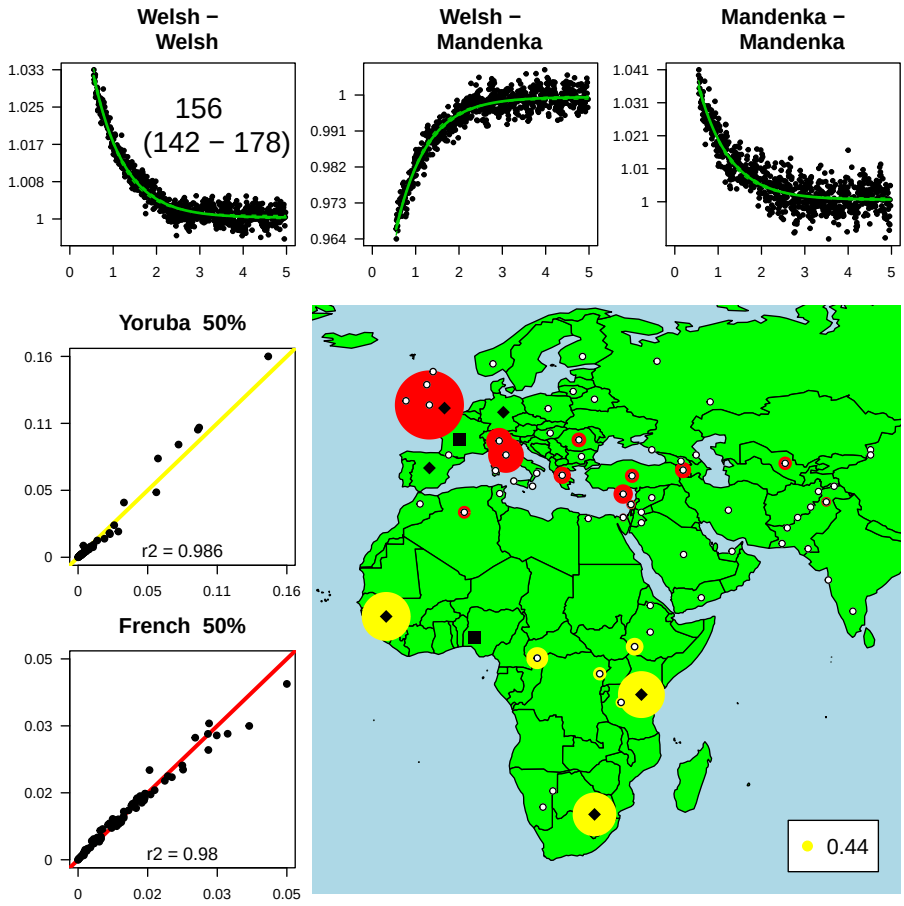



**Figure S5:** Results for simulations with 20 individuals admixed at 150 generations ago, with 50% of their DNA from Yoruba and 50% from French. Plot details as in Figure S3.

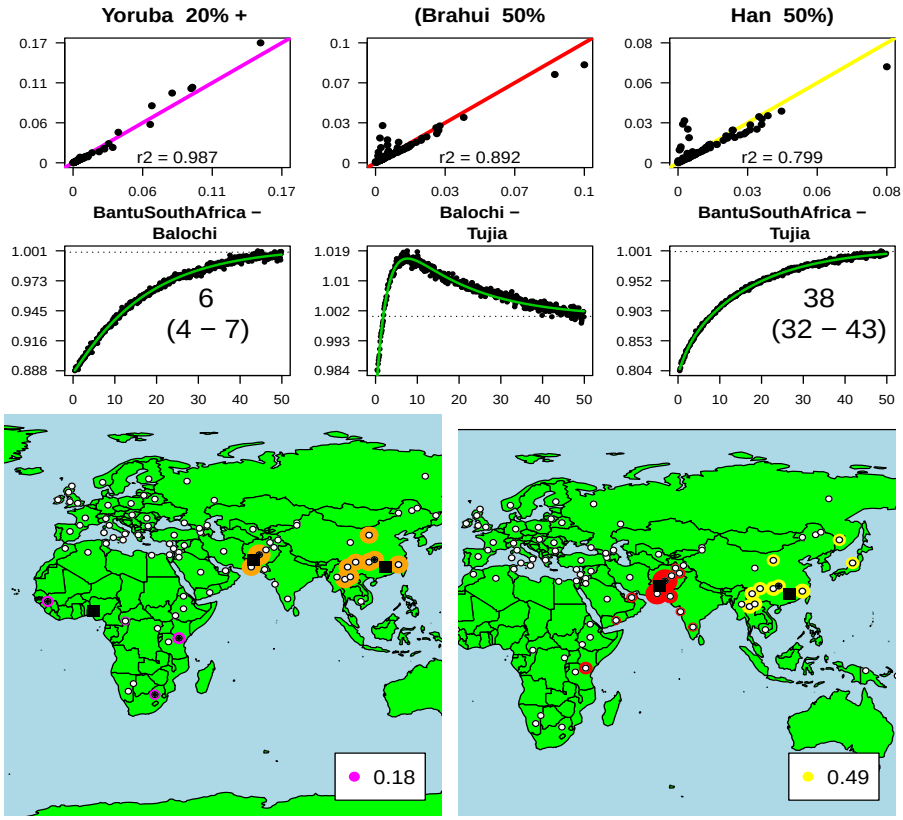



**Figure S6:** Results for simulations with 20 individuals admixed at 7 generations ago, with 20% of their DNA from Yoruba and 80% of their DNA from a group previously admixed 37 generations ago as 50% Brahui and 50% Han. Plot details as in Figure S3, but with three sources of admixture and two separate events. In the maps, circles depict donor populations inferred to represent each source, with sizes proportional to the differences between the two sources' inferred copying vectors (see Note Note S4.8 for details). The total estimated proportions from the purple and yellow sources are given in the legends.

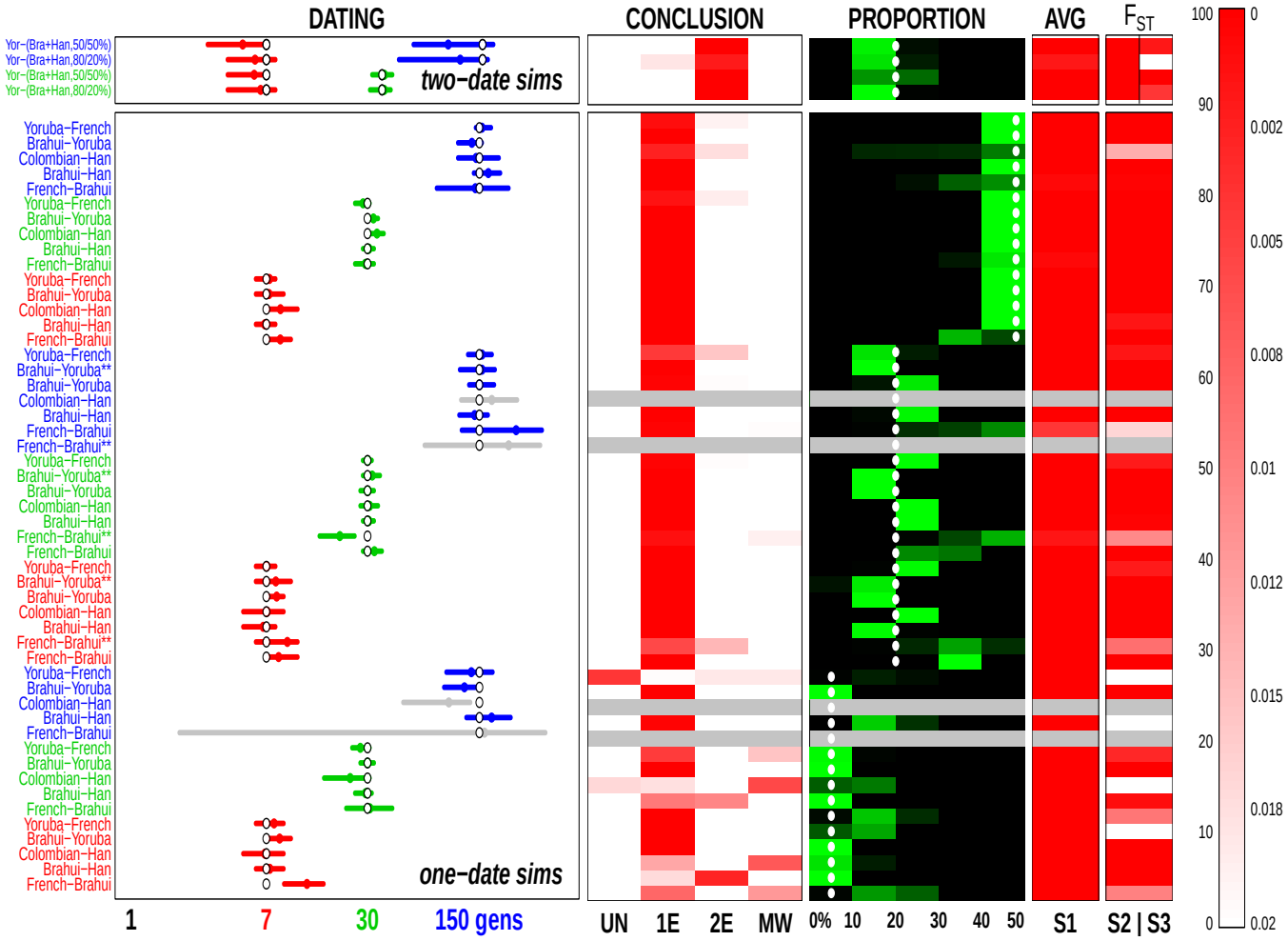



**Figure S7:** Performance of our approach in 55 "real-sample" simulations scenarios, involving different source groups (given at far left, with ** indicating "half-admixture" simulations), dates and/or proportions of admixture. First column ("Dating"): Inferred 95% bootstrap confidence intervals for admixture date(s) in generations (x-axis) colored by true admixture date (white dots denote truth). The top four rows have admixture simulated at two distinct times. Populations shown in grey here and in remaining columns are those where no significant evidence of admixture was detected by GLOBETROTTER ($p > 0.01$). Second column ("Conclusion"): Analysis of 101 replicate datasets generated under each of the 55 scenarios and analyzed separately. Redder colours indicate increasing fraction of cases GLOBETROTTER could not characterize admixture (UN), infers evidence ($p < 0.05$) of admixture at more than one time (2E, the correct choice for the upper 4 events), infers evidence of three admixing groups (MW), or did not reject a simple admixture event (1E, the correct choice for the lower 51 scenarios). Scale bar (0-100%) on right of figure. Third column ("Proportion"): Heatmap showing distribution of inferred admixture proportion (x-axis) for same scenarios for all bootstraps for which the inferred admixture was not "uncertain" (true proportion given with white dots), with green showing high density (for the most recent event in the top four cases). Fourth column ("AVG $F_{ST}$"): Mean $F_{ST}$ between the true source group and the single present-day group inferred to best represent this source, for same scenarios for which the inferred admixture was not "uncertain"; scale bar (0-0.02) shown on right of figure. The columns show $F_{ST}$ with each admixing source as ordered in the labels at far left of the figure; red indicates near exact match to true admixing source.

| | | | | | | | first event | | | | second event | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation Name** | **n** | $p$ | **UN** | **1D** | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **%** | **source 1** | **source 2** |
| **ONE DATE SIMULATIONS** | | | | | | | | | | | | | | |
| Brahui-Yoruba,7g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 8 (7-9) | 0.18-0.2 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,7g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 7 (6-9) | 0.47-0.5 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,7g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 7 (6-8) | 0.21-0.25 | 0-0 | 0.002-0.002 | – | – | – | – |
| Yoruba-French,7g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 7 (6-8) | 0.45-0.48 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Han,7g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 7 (5-8) | 0.19-0.2 | 0-0 | 0-0.001 | – | – | – | – |
| Brahui-Han,7g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 7 (6-8) | 0.49-0.5 | 0-0.003 | 0-0.003 | – | – | – | – |
| Brahui-Yoruba,7g,95/5% | 40 | <.01 | 0 | 1 | 0 | 0 | 8 (7-10) | 0.05-0.06 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,7g,95/5% | 40 | <.01 | 0 | 1 | 0 | 0 | 8 (6-9) | 0.08-0.14 | 0-0 | 0-0.152 | – | – | – | – |
| Brahui-Yoruba,7g,80/20%** | 40 | <.01 | 0 | 1 | 0 | 0 | 8 (6-10) | 0.1-0.14 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,30g,50/50% | 40 | <.01 | 0 | 0.99 | 0.01 | 0 | 33 (30-35) | 0.46-0.47 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Han,7g,95/5% | 40 | <.01 | 0 | 0.15 | 0.85 | 0 | 7 (6-9) | 0.06-0.07 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,30g,50/50% | 40 | <.01 | 0 | 0.92 | 0.08 | 0 | 28 (25-31) | 0.48-0.5 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,30g,95/5% | 40 | <.01 | 0 | 1 | 0 | 0 | 30 (27-33) | 0.06-0.06 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,30g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 30 (27-33) | 0.19-0.2 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Han,30g,80/20% | 40 | <.01 | 0 | 0.99 | 0.01 | 0 | 31 (28-33) | 0.21-0.22 | 0-0 | 0-0.001 | – | – | – | – |
| Brahui-Han,30g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 31 (28-33) | 0.48-0.5 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,30g,80/20% | 40 | <.01 | 0 | 0.98 | 0.02 | 0 | 30 (28-32) | 0.21-0.23 | 0-0 | 0.002-0.002 | – | – | – | – |
| Colombian-Han,7g,50/50% | 14 | <.01 | 0 | 1 | 0 | 0 | 9 (7-11) | 0.46-0.5 | 0-0 | 0-0 | – | – | – | – |
| Colombian-Han,7g,80/20% | 14 | <.01 | 0 | 1 | 0 | 0 | 7 (5-9) | 0.23-0.28 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,30g,95/5% | 40 | <.01 | 0 | 0.76 | 0 | 0.24 | 27 (24-31) | 0.06-0.11 | 0-0 | 0-0.006 | – | – | – | – |
| Brahui-Han,30g,95/5% | 40 | <.01 | 0 | 0.52 | 0.48 | 0 | 29 (25-32) | 0.06-0.07 | 0-0 | 0.001-0.001 | – | – | – | – |
| Brahui-Yoruba,30g,80/20%** | 40 | <.01 | 0 | 0.99 | 0.01 | 0 | 32 (28-36) | 0.12-0.14 | 0-0 | 0-0 | – | – | – | – |
| French-Brahui,7g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 9 (7-10) | 0.35-0.43 | 0-0 | 0-0 | – | – | – | – |
| Colombian-Han,30g,50/50% | 14 | <.01 | 0 | 1 | 0 | 0 | 34 (30-38) | 0.44-0.49 | 0-0 | 0-0 | – | – | – | – |
| French-Brahui,7g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 8 (7-11) | 0.32-0.37 | 0-0 | 0-0 | – | – | – | – |
| Colombian-Han,30g,80/20% | 14 | <.01 | 0 | 1 | 0 | 0 | 31 (27-35) | 0.23-0.28 | 0-0 | 0-0 | – | – | – | – |
| Colombian-Han,7g,95/5% | 14 | <.01 | 0 | 0.35 | 0 | 0.65 | 7 (5-9) | 0.07-0.12 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,150g,50/50%† | 40 | <.01 | 0 | 0.94 | 0.06 | 0 | 156 (142-178) | 0.44-0.5 | 0-0 | 0-0 | – | – | – | – |
| French-Brahui,30g,50/50% | 40 | <.01 | 0 | 1 | 0 | 0 | 29 (25-33) | 0.38-0.48 | 0-0.003 | 0-0 | – | – | – | – |
| Brahui-Yoruba,150g,50/50%† | 40 | <.01 | 0 | 1 | 0 | 0 | 135 (110-156) | 0.44-0.5 | 0-0 | 0-0 | – | – | – | – |
| Yoruba-French,150g,80/20%† | 40 | <.01 | 0 | 0.77 | 0.23 | 0 | 156 (127-180) | 0.16-0.22 | 0-0 | 0-0.002 | – | – | – | – |
| French-Brahui,7g,80/20%** | 40 | <.01 | 0 | 0.71 | 0.29 | 0 | 9 (6-11) | 0.24-0.47 | 0-0 | 0.009-0.009 | – | – | – | – |
| Brahui-Yoruba,150g,80/20%† | 40 | <.01 | 0 | 0.98 | 0.02 | 0 | 151 (129-186) | 0.19-0.25 | 0-0 | 0-0 | – | – | – | – |
| French-Brahui,30g,80/20% | 40 | <.01 | 0 | 1 | 0 | 0 | 33 (28-37) | 0.24-0.38 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Han,150g,80/20%† | 40 | <.01 | 0 | 1 | 0 | 0 | 140 (112-170) | 0.2-0.26 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,150g,95/5%† | 40 | <.01 | 0 | 1 | 0 | 0 | 121 (90-146) | 0.06-0.09 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Han,150g,50/50%† | 40 | <.01 | 0 | 1 | 0 | 0 | 171 (138-203) | 0.45-0.5 | 0-0 | 0-0 | – | – | – | – |
| Brahui-Yoruba,150g,80/20%**† | 40 | <.01 | 0 | 0.99 | 0.01 | 0 | 156 (113-188) | 0.12-0.18 | 0-0 | 0-0 | – | – | – | – |
| French-Brahui,7g,95/5% | 40 | <.01 | 0 | 0.59 | 0 | 0.41 | 13 (9-16) | 0.11-0.25 | 0-0 | 0.009-0.018 | – | – | – | – |
| French-Brahui,30g,80/20%** | 40 | <.01 | 0 | 0.93 | 0 | 0.07 | 20 (15-25) | 0.3-0.5 | 0-0.01 | 0.009-0.022 | – | – | – | – |
| Colombian-Han,150g,50/50%† | 14 | <.01 | 0 | 0.86 | 0.14 | 0 | 143 (110-199) | 0.13-0.49 | 0-0 | 0-0.135 | – | – | – | – |
| Brahui-Han,150g,95/5%† | 40 | <.01 | 0 | 0.98 | 0.01 | 0.01 | 179 (124-236) | 0.17-0.22 | 0-0 | 0-0.032 | – | – | – | – |
| French-Brahui,30g,95/5% | 40 | <.01 | 0 | 1 | 0 | 0 | 31 (22-43) | 0.1-0.27 | 0-0 | 0.009-0.018 | – | – | – | – |
| French-Brahui,150g,50/50%† | 40 | <.01 | 0 | 1 | 0 | 0 | 142 (81-229) | 0.26-0.48 | 0-0.003 | 0-0.003 | – | – | – | – |
| French-Brahui,150g,80/20%† | 40 | <.01 | 0 | 0.98 | 0 | 0.02 | 254 (116-370) | 0.22-0.49 | 0-0.01 | 0-0.022 | – | – | – | – |
| Colombian-Han,30g,95/5% | 14 | <.01 | 0.16 | 0.13 | 0 | 0.71 | 23 (16-31) | 0.07-0.16 | 0-0 | 0-0.135 | – | – | – | – |

Continued on next page

| Table S1 – continued from previous page | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | first event | | | | second event | | | |
| **Simulation Name** | **n** | $p$ | **UN** | **1D** | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **%** | **source 1** | **source 2** |
| Yoruba-French,150g,95/5%† | 40 | <.01 | 0.78 | 0.02 | 0.1 | 0.1 | 134 (93-182) | 0.09-0.36 | 0-0 | 0.152-0.153 | – | – | – | – |
| Colombian-Han,150g,80/20%† | 14 | 0.05 | – | – | – | – | – | – | – | – | – | – | – | – |
| French-Brahui,150g,80/20%**† | 40 | 0.11 | – | – | – | – | – | – | – | – | – | – | – | – |
| Colombian-Han,150g,95/5%† | 14 | 0.25 | – | – | – | – | – | – | – | – | – | – | – | – |
| French-Brahui,150g,95/5%† | 40 | 0.36 | – | – | – | – | – | – | – | – | – | – | – | – |
| **TWO DATE SIMULATIONS** | | | | | | | | | | | | | | |
| Yor,7g,20%-(Br+Hn,157g,80/20%)‡ | 40 | <.01 | 0 | 0.11 | 0.89 | 0 | 6 (4-8) | 0.18-0.23 | 0-0 | – | 114 (47-170) | 0.18-0.47 | 0-0 | 0-0.187 |
| Yor,7g,20%-(Br+Hn,157g,50/50%)‡ | 40 | <.01 | 0 | 0.01 | 0.99 | 0 | 5 (3-7) | 0.19-0.21 | 0-0 | – | 96 (58-184) | 0.28-0.5 | 0-0 | 0-0 |
| Yor,7g,20%-(Br+Hn,37g,80/20%) | 40 | <.01 | 0 | 0 | 1 | 0 | 6 (4-8) | 0.18-0.2 | 0-0 | – | 38 (31-42) | 0.21-0.26 | 0-0 | 0-0.001 |
| Yor,7g,20%-(Br+Hn,37g,50/50%) | 40 | <.01 | 0 | 0 | 1 | 0 | 6 (4-7) | 0.18-0.21 | 0-0 | – | 38 (32-43) | 0.48-0.5 | 0-0 | 0-0 |
| **NO-ADMIXTURE SIMULATIONS** | | | | | | | | | | | | | | |
| Han-Japanese,1000g,50/50% | 40 | 0.29 | – | – | – | – | – | – | – | – | – | – | – | – |
| Sindhi-Pathan,1000g,50/50% | 40 | 0.75 | – | – | – | – | – | – | – | – | – | – | – | – |
| Brahui-Brahui,1000g,50/50% | 40 | 0.77 | – | – | – | – | – | – | – | – | – | – | – | – |
| French-French,1000g,50/50% | 40 | 0.82 | – | – | – | – | – | – | – | – | – | – | – | – |
| Basque-French,1000g,50/50% | 40 | 0.86 | – | – | – | – | – | – | – | – | – | – | – | – |

**Table S1:** Power results for all inferred events in the "real-sample" simulations (**indicates "half-admixture simulations"). Columns give inferred dates (+ 95% CI) across 101 bootstrap replicates, plus inferred 95% range of admixture proportions ("%") from the minority source and 95% range of $F_{ST}$ between the single present-day population that best matches each admixing source group and the true source group ("source 1" and "source 2") across all 101 bootstrap replicates for which inferrence was not "uncertain". For all except the "TWO DATE SIMULATIONS", "source1" and "source2" refer to the majority and minority admixing source populations, respectively. For "TWO DATE SIMULATIONS", "source1" and "source2" in the "first event" columns refer to the minority and majority sources, respectively, in the recent event (i.e. with "source1" the Yoruba) and "source1" and "source2" in the "second event" columns refer to the majority and minority sources, respectively, in the older event. When the true proportion of admixture is 50%, the minority and majority sources are arbitrary; in such cases, we assign inferred sources based on which gives the lowest combined $F_{ST}$ with the truth. $n$ gives the number of individuals per simulation; $p$ gives the $p$-value for concluding any admixture; "UN", "1D", "2E", "MW" give the proportion of 101 bootstraps for which the admixture conclusion is "uncertain", "one date", "multiple dates", and "one date, multiway" admixture, respectively, as described in Note S4.6. †Dates for these simulations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these simulations were inferred using the "multiple-date" grid; see Note S4.3.1.

## S5.3 Comparison to ROLLOFF using "real-sample" simulations

Recently developed software called ROLLOFF, first described in (4) and released as part of a suite of software called ADMIXTOOLS (27), also aims to date instantaneous admixture events. ROLLOFF uses a different approach, based on allele frequency differences at individual SNPs, to distinguish between ancestral sources. At present, ROLLOFF considers the case of 2-way admixture at a single time in history, using user-input reference populations to provide information on the two sources.

To assess the benefits of using haplotype information, rather than individual SNPs as in ROLLOFF, for inferring dates, we re-analyzed all 45 of our "one-date simulations" (see Note S5.1.1) in a manner that matches the protocol a user would typically follow when using ROLLOFF. We note that our method differs in several ways to ROLLOFF. In particular our approach aims to infer source populations, and considers admixture at multiple times or between multiple groups. These differences limit the range of scenarios in which comparison is possible. To perform the fairest comparison, we ran both our approach and ROLLOFF restricting to only two sets of "surrogate populations" as donors to describe the admixture event, selecting surrogates that closely match the true admixing source groups genetically. The sampled populations we used as the surrogates for each simulation are given in Table S2.

| **simulation sources** | **source 1 surrogate pops** | **source 2 surrogate pops** |
|---|---|---|
| Brahui, Yoruba | Balochi (21) | Mandenka (22) |
| Brahui, Han | Balochi (21) | HanNchina, Tujia (20) |
| Colombian, Han | Maya (21) | HanNchina, Tujia (20) |
| French, Brahui | English, Ireland, Scottish, Welsh (23) | Balochi (21) |
| Yoruba, French | Mandenka (22) | English, Ireland, Scottish, Welsh (23) |

**Table S2:** The list of sampled populations used to represent each of the two ancestral admixing sources in simulations for our ROLLOFF comparison analysis. The first column gives the true admixing sources and the second and third columns list the populations that comprise the surrogates for source one and source two, respectively, in our analysis of Note S5.3. For each surrogate population, all samples available after quality control were used (the total number of individuals for each surrogate is given in parentheses).

For this comparison, we kept the parameters of our approach and ROLLOFF as similar as possible, for example using the same grid bin-size (0.1cM, as recommended by ROLLOFF) and fitting the curve from 1 to 50cM. Otherwise for our approach we used the same analysis protocol outlined in Note S4. For any parameters ROLLOFF requires that our analysis does not, we used default values in the ROLLOFF software.

In Figure S8, for several of the simulated populations we compare our coancestry curves (to which dates were fitted) to the analogous from ROLLOFF. The curves generated under our approach are much more tightly clustered around the truth in cases where the source populations are genetically similar, such as the simulations with French admixing with Brahui and Colombian admixing with Han. Comparisons of date estimation for all 45 simulations are given in the top panels of Figures S9 and S10, which illustrate our model's increased accuracy and precision, respectively, over ROLLOFF. For example among all 42 simulations for which our model concludes the presence of admixture, four have a date point estimate off by at least 25% from the true value, compared to ten for ROLLOFF (Figure S9, top). Furthermore the standard errors of our estimates are smaller than those of ROLLOFF in 38 of these 41 simulations (Figure S10, top; for one of the 42 simulations, ROLLOFF failed with an error when trying to infer the standard error), reflecting the decreased variability in our coancestry curves of Figure S8.

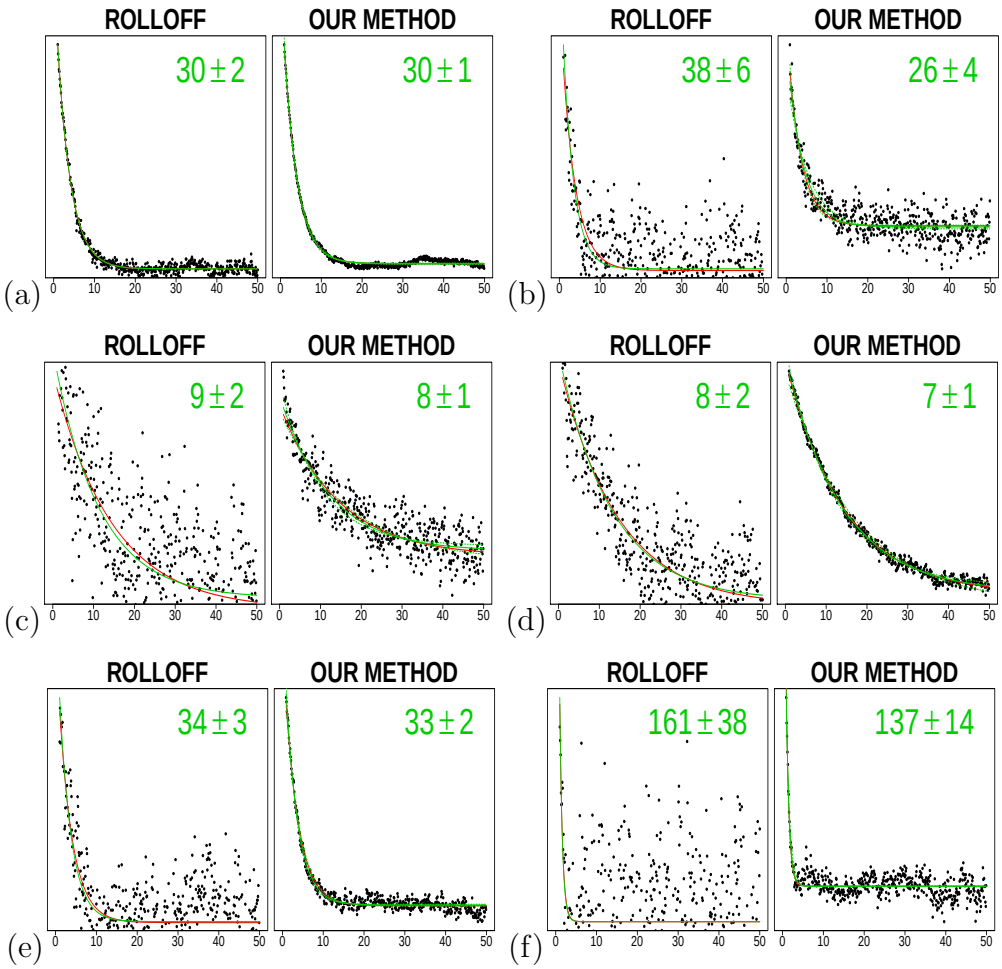



**Figure S8:** Coancestry curves for ROLLOFF (left) versus our method (right) for the following simulated admixture events: (a) 80% contribution from Brahui plus 20% contribution from Yoruba, 30 generations ago; (b) 50% French + 50% Brahui, 30 gen ago; (c) 80% French + 20% Brahui, 7 gen ago; (d) 80% Colombian + 20% Han, 7 gen ago; (e) 80% Colombian + 20% Han, 30 gen ago; (f) 50% Colombian + 50% Han, 150 gen ago. Inferred dates and standard errors are given at top right in each plot. Green lines give exponential decay curves with rates equal to inferred dates of admixture; red lines give exponential decay curves with rates equal to true dates of admixture. For our results, we show the coancestry curve for the surrogate group representing the minority admixing source. To calculate standard errors, for ROLLOFF we used the Weighted Block Jackknife procedure described in (50) as recommended by the authors and for our method took the standard deviation across 100 bootstrap re-samples.

To compare ROLLOFF to our recommended manner of using our model, in the bottom of Figures S9 and S10 we compare the ROLLOFF date estimates to those of the analysis we report in Note S5.2 that used the protocol outlined in Note S4, i.e. for which we did not pre-select surrogate populations but instead used all available donors to infer source groups. In this analysis, we again conclude 42 simulations to have detectable admixture. Only three among these 42 have date point estimates off by at least 25% of the true value, compared to ten for ROLLOFF (Figure S9, bottom). Our standard errors are smaller than that of ROLLOFF in 37 of these 41 simulations (Figure S10, bottom; again note that ROLLOFF failed to infer a standard error in one of the 42 simulations).

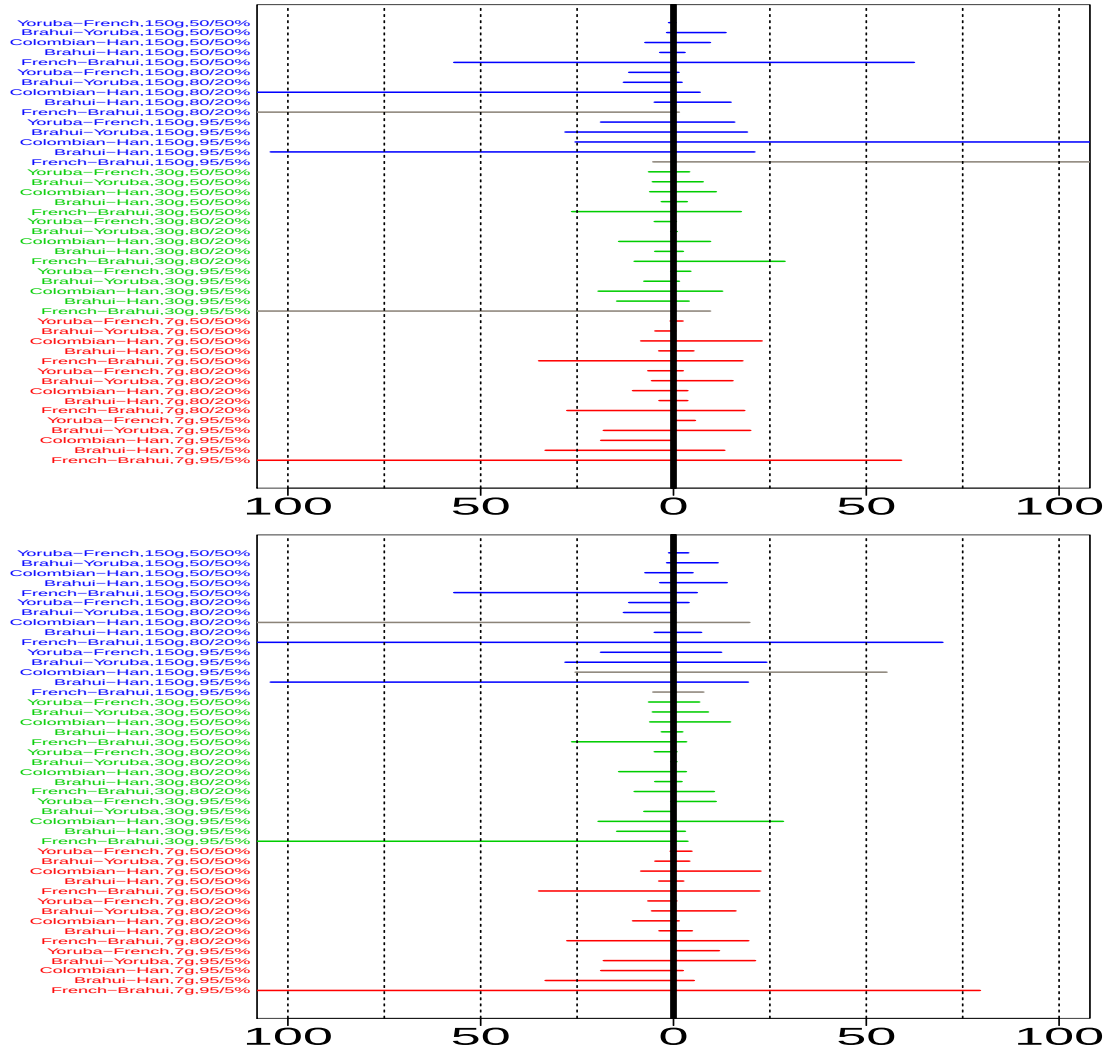



**Figure S9:** Comparison between our date point estimates (top = using donors from Table S2; bottom = using all available donors) and those of ROLLOFF among 45 populations simulated with a single admixture event. In each row, lines stretch to the percentage by which the date point estimate from ROLLOFF (left) and our method (right) differs from the true date of admixture. The simulated populations (rows) are labeled along the y-axis according to their two admixing source populations and the dates and proportions of admixture, ordered by date of admixture followed by proportion of admixture followed by $F_{ST}$ between the source populations. These labels and the lines are colored according to the true simulated date of admixture (red=7 generations, green=30, blue=150). Lines in grey denote populations for which our model concluded "no admixture".

## S5.4 Details of "coalescent-based" simulations

We also tested our model using the approximate coalescence simulation software MaCs (14), simulating 11 populations with the history depicted in Figure S11. While it is impossible to simulate a scenario that perfectly captures the history of modern human groups, with references in the literature on appropriate split times and population sizes often uncertain or disagreeing, we tried to capture major features of world-wide human migrations and splits, informing our scenario using a number of recent publications (15; 16; 17; 18; 51). For example, populations 1-4 are meant to mimic genetic diversity among African populations (15). The split at 2500 generations ago and subsequent bottleneck in populations 5-11 mimics the "out-of-Africa" event (15; 16; 51), and the split between population set 4-7 and population set 8-11 at 1000 generations ago mimics the split between Western Eurasia populations and East Asian populations (18; 17),

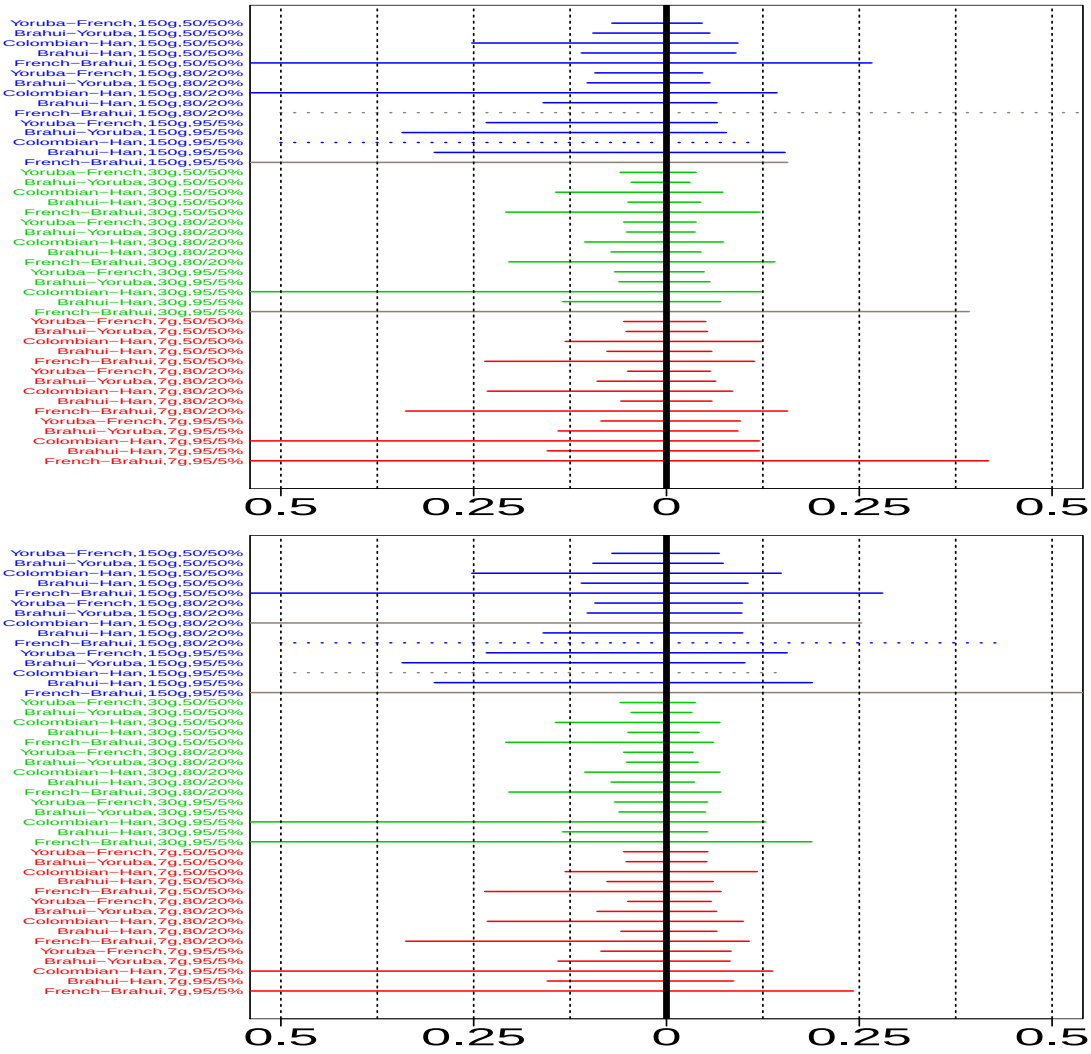



**Figure S10:** Comparison between the standard errors of our date estimates (top = using donors from Table S2; bottom = using all available donors) and those of ROLLOFF among 45 populations simulated with a single admixture event. In each row, lines stretch to the estimated standard error of the date estimates for ROLLOFF (left) and our method (right) divided by the true date of admixture. The simulated populations (rows) are ordered as in Figure S9 and colored (as with lines) by the true date of admixture (red=7 generations, green=30, blue=150). Lines in grey denote populations for which our model concluded "no admixture"; the dotted grey lines indicate that in addition ROLLOFF failed to finish without error for these populations. To calculate standard errors, for ROLLOFF we used the Weighted Block Jackknife procedure described in (50), removing one chromosome at a time as recommended by the authors, and for our method took the standard deviation across 100 bootstrap re-samples.

with a subsequent bottleneck in the latter. We also included continuous symmetric migration between populations 9 and 10 and between populations 10 and 11 such that each population's fraction of new migrants increased by 0.00025 each generation for 375 generations (continuing until present-day), in order to test our model's robustness to a scenario of stable long-term migration between nearby groups (so that 18-19% of current population 10's haplotypes are migrants, and corresponding to 10 migrant haplotypes per generation into population 10).

We simulated 22 chromosomes, with lengths and variation in recombination rates for each based on HapMap Phase 2 build 36 genetic maps (www.hapmap.org). We assumed a genome-wide average recombination rate of $1.25 \times 10^{-8}$ per basepair and a mutation rate of $2.5 \times 10^{-8}$ per basepair, and sampled 200 haplotypes from each of populations 1-11. For each chromosome, we subsequently selected the number of SNPs that matched our real data (474,491 SNPs in total), selecting SNPs such that the minor allele frequency spectrum of our simulated dataset (across all populations) matched that of our real dataset (across all populations) across 100 equally spaced bins from 0 to 0.5. (We randomly selected SNPs to fill bins where we did not have enough such candidate SNPs.) Among our 200 sampled haplotypes per population, we used 150 of these to simulate admixed populations and the remaining 50 as sampled data in our analyses as described below. Using these 50 samples per population, we provide pairwise $F_{ST}$ values among all 11 populations in Table S4, which reinforces that our simulation scenario provides results similar to relationships among present-day human DNA collections. We also assessed whether linkage disequilibrium patterns in our simulated data matched that observed in our real dataset, by plotting pairwise squared correlation ($r^2$) values versus distance for 1000 randomly selected SNPs on chromosome 22. We found that the patterns of these curves for simulated populations 4, 5 and 8 were close matches to that observed in the Yoruba, French and Han populations, respectively (results omitted).

We next simulated six admixed populations, labeled A-F, that consisted of 50 haplotypes apiece, as shown in Table S3, using the 150 simulated haplotypes from populations 1-11 described above and the simulation technique described in Note S5.1 (with ancestry segments sampled so that admixed individuals never shared the same source haplotype at any location of the genome). Pairwise $F_{ST}$ values across these admixed populations and unadmixed populations 1-11, computed using the 50 haplotypes from each used in the analyses of Note S5.5 below, are provided in Table S4.

| **Simulation** | **Scenario** | **Date(s)** |
|---|---|---|
| **PopA** | 50% Pop5 + 45% Pop7 + 5% Pop8 | 36 |
| **PopB** | 95% Pop7 + 5% Pop8 | 30 |
| **PopC** | 95% Pop6 + 5% Pop3 [95% Pop6 + 5% Pop3] | 8 [58] |
| **PopD** | 97% Pop5 + 3% Pop4 | 21 |
| **PopE** | 70% Pop1 + 30% Pop5 | 60 |
| **PopF** | 75% Pop1 + 25% Pop3 | 100 |

**Table S3:** Details of all admixed populations generated using populations 1-11 (see Figure S11) of the "coalescent-based" simulations. Dates are given in generations from present. "Scenario" provides the proportions of admixture from each source population 1-11 used to simulate admixture. For PopC, which was simulated with two distinct dates of admixture, details of the second older event are given in brackets.

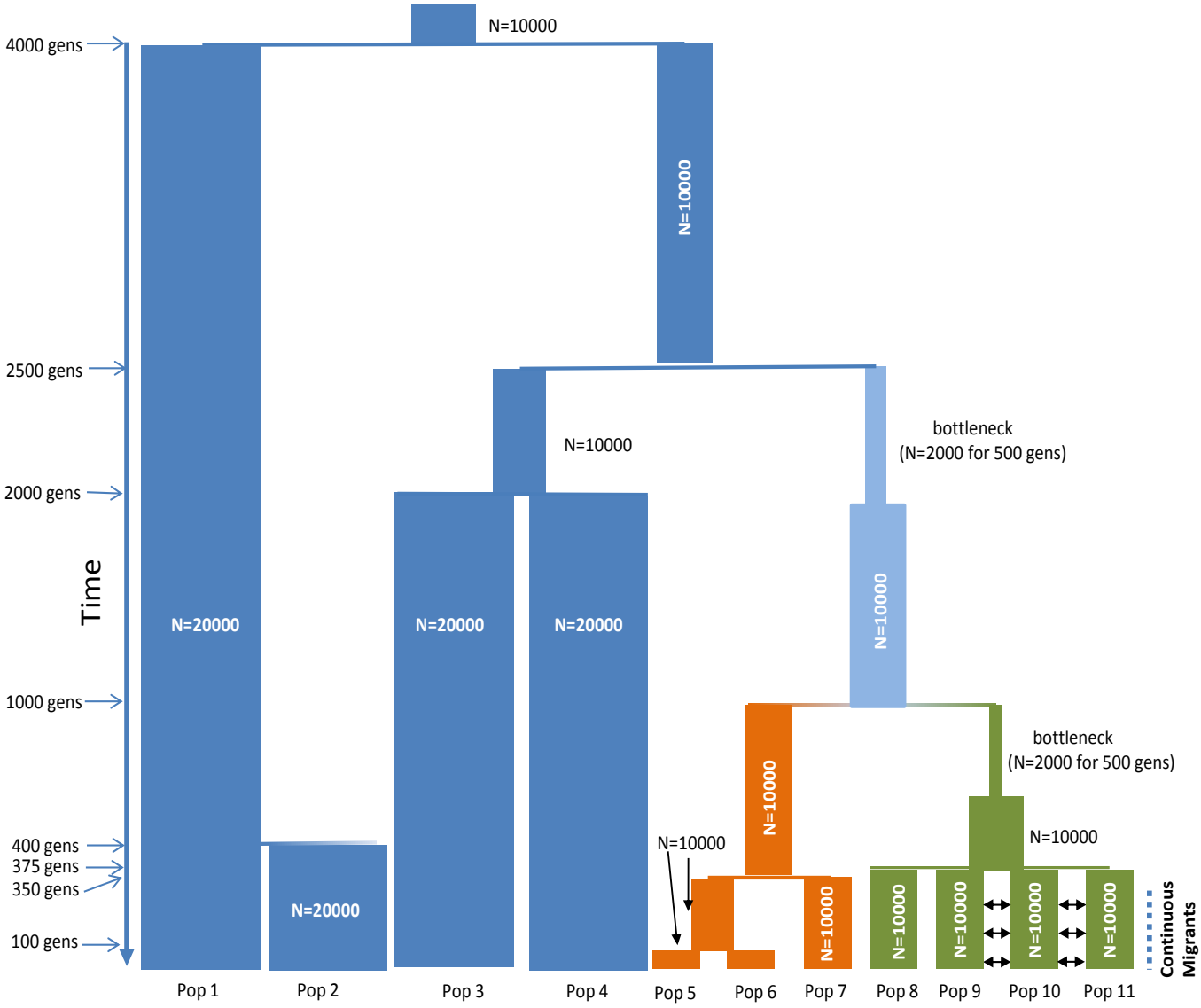



**Figure S11:** Simulated history for populations 1-11, generated using the coalescent-based software MaCS (14). Roughly speaking, populations 1-4 in blue are meant to represent diversity in African groups, with populations 5-7 in orange and 8-11 in green representing Western Eurasian and East Asian groups, respectively. 100 generations (gens) denotes the split between Pop5 and Pop6; 350 gens the split between Pop7 and Pop5/Pop6; 375 gens the simultaneous split of populations 8-11; 400 gens the split of Pop1 and Pop2; 1000 gens the split of Pop5/Pop6/Pop7 and Pop8/Pop9/Pop10/Pop11; 2000 gens the split of Pop3 and Pop4; 2500 gens the split of Pop3/Pop4 and Pop5/Pop6/Pop7/Pop8/Pop9/Pop10/Pop11; and 4000 gens the split of Pop1/Pop2 and all other populations.

| | **Pop1** | **Pop2** | **Pop3** | **Pop4** | **Pop5** | **Pop6** | **Pop7** | **Pop8** | **Pop9** | **Pop10** | **Pop11** | **PopA** | **PopB** | **PopC** | **PopD** | **PopE** | **PopF** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Pop1** | 0 | 0.01 | 0.115 | 0.114 | 0.185 | 0.185 | 0.186 | 0.228 | 0.229 | 0.226 | 0.227 | 0.178 | 0.182 | 0.168 | 0.179 | 0.02 | 0.008 |
| **Pop2** | 0.01 | 0 | 0.115 | 0.114 | 0.185 | 0.185 | 0.186 | 0.228 | 0.229 | 0.226 | 0.227 | 0.179 | 0.183 | 0.167 | 0.179 | 0.026 | 0.015 |
| **Pop3** | 0.115 | 0.115 | 0 | 0.048 | 0.136 | 0.136 | 0.136 | 0.18 | 0.181 | 0.177 | 0.178 | 0.129 | 0.133 | 0.113 | 0.13 | 0.079 | 0.066 |
| **Pop4** | 0.114 | 0.114 | 0.048 | 0 | 0.134 | 0.135 | 0.135 | 0.178 | 0.179 | 0.176 | 0.176 | 0.128 | 0.132 | 0.116 | 0.128 | 0.078 | 0.076 |
| **Pop5** | 0.185 | 0.185 | 0.136 | 0.134 | 0 | 0.005 | 0.018 | 0.098 | 0.099 | 0.096 | 0.096 | 0.004 | 0.016 | 0.006 | 0 | 0.094 | 0.153 |
| **Pop6** | 0.185 | 0.185 | 0.136 | 0.135 | 0.005 | 0 | 0.018 | 0.098 | 0.099 | 0.096 | 0.097 | 0.007 | 0.016 | 0.001 | 0.005 | 0.095 | 0.154 |
| **Pop7** | 0.186 | 0.186 | 0.136 | 0.135 | 0.018 | 0.018 | 0 | 0.098 | 0.098 | 0.095 | 0.096 | 0.005 | 0.001 | 0.017 | 0.017 | 0.099 | 0.154 |
| **Pop8** | 0.228 | 0.228 | 0.18 | 0.178 | 0.098 | 0.098 | 0.098 | 0 | 0.018 | 0.017 | 0.018 | 0.081 | 0.088 | 0.094 | 0.097 | 0.152 | 0.196 |
| **Pop9** | 0.229 | 0.229 | 0.181 | 0.179 | 0.099 | 0.099 | 0.098 | 0.018 | 0 | 0.016 | 0.018 | 0.083 | 0.09 | 0.094 | 0.097 | 0.153 | 0.197 |
| **Pop10** | 0.226 | 0.226 | 0.177 | 0.176 | 0.096 | 0.096 | 0.095 | 0.017 | 0.016 | 0 | 0.014 | 0.079 | 0.087 | 0.092 | 0.094 | 0.15 | 0.195 |
| **Pop11** | 0.227 | 0.227 | 0.178 | 0.176 | 0.096 | 0.097 | 0.096 | 0.018 | 0.018 | 0.014 | 0 | 0.08 | 0.087 | 0.092 | 0.095 | 0.151 | 0.195 |
| **PopA** | 0.178 | 0.179 | 0.129 | 0.128 | 0.004 | 0.007 | 0.005 | 0.081 | 0.083 | 0.079 | 0.08 | 0 | 0.004 | 0.007 | 0.004 | 0.09 | 0.147 |
| **PopB** | 0.182 | 0.183 | 0.133 | 0.132 | 0.016 | 0.016 | 0.001 | 0.088 | 0.09 | 0.087 | 0.087 | 0.004 | 0 | 0.016 | 0.016 | 0.096 | 0.151 |
| **PopC** | 0.168 | 0.167 | 0.113 | 0.116 | 0.006 | 0.001 | 0.017 | 0.094 | 0.094 | 0.092 | 0.092 | 0.007 | 0.016 | 0 | 0.005 | 0.082 | 0.134 |
| **PopD** | 0.179 | 0.179 | 0.13 | 0.128 | 0 | 0.005 | 0.017 | 0.097 | 0.097 | 0.094 | 0.095 | 0.004 | 0.016 | 0.005 | 0 | 0.089 | 0.148 |
| **PopE** | 0.02 | 0.026 | 0.079 | 0.078 | 0.094 | 0.095 | 0.099 | 0.152 | 0.153 | 0.15 | 0.151 | 0.09 | 0.096 | 0.082 | 0.089 | 0 | 0.014 |
| **PopF** | 0.008 | 0.015 | 0.066 | 0.076 | 0.153 | 0.154 | 0.154 | 0.196 | 0.197 | 0.195 | 0.195 | 0.147 | 0.151 | 0.134 | 0.148 | 0.014 | 0 |

**Table S4:** Pairwise $F_{ST}$ (49) values among all populations in the "coalescent-based" simulations.

## S5.5 Results of "coalescent-based" simulations

We performed two separate analyses of these "coalescent-based" simulations:

1. "full analysis" – we included 50 haplotypes from each of populations 1-11 and A-F
2. "hard analysis" – we included 50 haplotypes from each of populations 2, 4, 9-11 and A-F

The "full analysis" marks an ideal scenario where we have sampled haplotypes from each population involved in the admixture of populations A-F, though note that we do not include the exact parental haplotypes involved in the admixture for our analysis. The "hard analysis" marks a more difficult scenario where we have failed to sample several populations relevant to the admixture events, indeed including only one (Pop4) directly used to simulate events in populations A-F. In particular we have no good surrogate for our "Western Eurasian" populations 5-7 and will thus have to rely on admixed populations A-F to identify and describe many of these events.

In each analysis, we treat all individuals with the same population label as a single group, and repeat the analysis procedure described in Note S4, testing for evidence of admixture in each included group while using all others as surrogates. Our results are provided in Table S5. In both the "full" and "hard" analyses, we infer admixture in all (100%) of groups A-F ($p$-value $< 0.01$) and never identify a false event in populations 1-11 ($p$-value $> 0.1$). We correctly conclude that PopC has multiple dates in both analyses. We correctly conclude that PopA has multiway admixture in the "full" analysis and just miss the $p$-value threshold of 0.05 for concluding multiway admixture in the "hard" analysis ($p$-value $= 0.06$). We nonetheless very accurately describe the strongest signal of admixture in the latter case.

In the "full" analysis, GLOBETROTTER performs particularly well at describing the inferred admixture events. The 95% CI of our date estimates contain the truth in all admixed populations. With the exception of PopC, our proportion estimates for the recent admixture are within 5% of the truth, and GLOBETROTTER always selects the correct sampled population to represent the admixing source group.

In the "hard" analysis, the 95% CI of our date estimates contains the truth in all admixed populations but PopB, for which the truth is 30 generations and our model infers 21-28 generations. Despite our deliberately highly incomplete sampling of the true admixing source populations (for 8 out of 13 true sources, $F_{ST} \geq 0.048$ to any sampled unadmixed group), our proportion and source estimation is generally accurate. (Note that though Pop1, Pop3 and Pops 5-8 were not used as surrogates in this analysis, we still inferred their copying vectors by allowing them to copy from the "hard" analysis populations only using CHROMOPAINTER. Therefore our inferred copying vector for each source was allowed to have a "best-match" with these unsampled populations.) Proportions were within 5% of the truth and $F_{ST} < 0.006$ to the true sources, with $F_{ST} = 0$ in five cases, including the only case where the true source was sampled. The exceptions to this are PopC and PopF. In both cases, the minority admixing group is unsampled Pop3, which has no sampled surrogate in the dataset (the most similar sampled group is Pop4, with a simulated separation time of over 50,000 years). In these two populations, the date estimation and description of the majority admixing source groups are very accurate, suggesting this is robust to no surrogate existing for the minority source group. Further, the minority source identified for PopF is Pop4, the best sampled surrogate for Pop3, and the minority source identified for PopC is the second best sampled surrogate for Pop3 (PopF, $F_{ST} = 0.066$, with 25% ancestry from Pop3), implying some information is still captured about the true underlying source in these difficult cases.

| | | | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation** | **true date** | **true %** | $p$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** | **date** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** |
| **FULL ANALYSIS** | | | | | | | | | | | | | | |
| Pop1 | – | – | 0.84 | – | – | – | – | – | – | – | – | – | – | – |
| Pop2 | – | – | 0.76 | – | – | – | – | – | – | – | – | – | – | – |
| Pop3 | – | – | 0.59 | – | – | – | – | – | – | – | – | – | – | – |
| Pop4 | – | – | 0.11 | – | – | – | – | – | – | – | – | – | – | – |
| Pop5 | – | – | 1 | – | – | – | – | – | – | – | – | – | – | – |
| Pop6 | – | – | 0.3 | – | – | – | – | – | – | – | – | – | – | – |
| Pop7 | – | – | 1 | – | – | – | – | – | – | – | – | – | – | – |
| Pop8 | – | – | 0.93 | – | – | – | – | – | – | – | – | – | – | – |
| Pop9 | – | – | 0.71 | – | – | – | – | – | – | – | – | – | – | – |
| Pop10 | – | – | 0.42 | – | – | – | – | – | – | – | – | – | – | – |
| Pop11 | – | – | 0.61 | – | – | – | – | – | – | – | – | – | – | – |
| PopA$^M$ | 36 | 45 | <.01 | 0.992 | 1 | 0.71 | <.01 | 36 (32-38) | 40 | PopB; 0.001 [Pop7; 0] | Pop5; 0 [Pop5; 0] | – | Pop8; 0 | Pop7; 0 |
| PopB$^1$ | 30 | 5 | <.01 | 0.989 | 1 | – | 0.78 | 28 (25-32) | 8 | Pop8; 0 | Pop7; 0 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | <.01 | 0.982 | 1 | 0.02 | 0.8 | 8 (3-11) | 18 | Pop3; 0 [Pop3; 0] | Pop6; 0 [Pop6; 0] | 65 (33-84) | Pop3; 0 | Pop6; 0 |
| PopD$^1$ | 21 | 3 | <.01 | 0.996 | 1 | – | 0.82 | 21 (18-23) | 5 | Pop4; 0 | Pop5; 0 | – | – | – |
| PopE$^1$† | 60 | 30 | <.01 | 0.993 | 1 | 0.86 | 0.97 | 64 (60-69) | 34 | Pop5; 0 | Pop1; 0 | – | – | – |
| PopF$^1$† | 100 | 25 | <.01 | 0.979 | 1 | 0.92 | 0.97 | 96 (85-106) | 27 | Pop3; 0 | Pop1; 0 | – | – | – |
| **HARD ANALYSIS** | | | | | | | | | | | | | | |
| Pop2 | – | – | 0.36 | – | – | – | – | – | – | – | – | – | – | – |
| Pop4 | – | – | 0.43 | – | – | – | – | – | – | – | – | – | – | – |
| Pop9 | – | – | 0.72 | – | – | – | – | – | – | – | – | – | – | – |
| Pop10 | – | – | 0.38 | – | – | – | – | – | – | – | – | – | – | – |
| Pop11 | – | – | 0.36 | – | – | – | – | – | – | – | – | – | – | – |
| PopA$^1$ | 36 | 45 | <.01 | 0.992 | 1 | 0.98 | 0.06 | 33 (30-36) | 41 | PopB; 0.001 | PopD; 0 | – | – | – |
| PopB$^1$ | 30 | 5 | <.01 | 0.978 | 1 | – | 0.33 | 25 (21-28) | 7 | Pop8; 0 | PopA; 0.005 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | <.01 | 0.981 | 1 | 0.02 | 0.39 | 8 (5-10) | 16 | Pop8; 0.18 [PopF; 0.066] | Pop6; 0 [PopD; 0.005] | 68 (44-86) | PopF; 0.066 | PopD; 0.005 |
| PopD$^1$ | 21 | 3 | <.01 | 0.996 | 1 | 0.96 | 0.45 | 21 (18-25) | 8 | Pop4; 0 | Pop6; 0.005 | – | – | – |
| PopE$^1$† | 60 | 30 | <.01 | 0.995 | 1 | 1 | 0.97 | 61 (56-65) | 35 | PopD; 0 | Pop1; 0 | – | – | – |
| PopF$^1$† | 100 | 25 | <.01 | 0.947 | 1 | 0.76 | 0.73 | 99 (87-115) | 37 | Pop4; 0.048 | Pop2; 0.01 | – | – | – |

**Table S5:** Inferred dates (+ 95% CI; in generations from present), proportions of admixture from minority source 1 (%) and the single population that best matches each admixing source group (as well as the $F_{ST}$ – as given in Table S4 – between this "best match" population and the true source), for all inferred events in our analysis of the "coalescent-based" simulations. $p$ gives the $p$-value for concluding any admixture; $R_1 = \max_i R_1^i$ as described in Note S4.8; "2E" and "MW" give $p$-values for multiple-dates and multiway admixture, respectively, as described in Note S4.6; and $FQ_B \equiv \min(FQ_1 + FQ_2, FQ_1^{\text{NULL}} + FQ_2^{\text{NULL}})$ is defined in Notes S4.5,S4.6 and used for calling events as "uncertain". Populations with $p = 1$ (i.e. Pop5, Pop7) failed to run without error in the initial proportion estimation step, implying a single inferred donor contributing to the mixture representation, and so implying no detected admixture. For the first event in populations with MW or 2E $< 0.05$, for each source we provide both the best match and (in brackets) the population most represented in the inferred copying vector relative to the other source. For the second event in these cases, we only provide the population most represented in the inferred copying vector relative to the other source. The "true %" column refers to the true proportion of admixture in the most recent or strongest (i.e. "first") event. Our model's inferred conclusion for each population is given next to its name in the first column, with $^1$= "one-date", $^M$= "one date, multiway", and $^2$= "multiple-dates". †Dates/proportions/sources for these populations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these simulations were inferred using the "multiple-date" grid; see Note S4.3.1. See Table S3 for details of the admixed populations PopA-PopF.

### S5.5.1 robustness to fewer sampled individuals

For the "hard analysis", we tested the effect of having fewer than 50 haplotypes in each of the admixed populations A-F. For each of these populations, we made new coancestry curves and copying vectors that used only 10 or 20 haplotypes from the given population, and then re-inferred dates, proportions, and admixing source groups. The results are provided in Table S6. Performance is very similar to the "hard analysis" results of Table S5.

| | | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation** | **true date** | **true %** | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** | **date** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** |
| **10 HAPLOTYPES** | | | | | | | | | | | | | |
| PopA$^M$ | 36 | 45 | 0.956 | 1 | 0.82 | 0.02 | 34 | 0.39 | PopB; 0.001 [PopB; 0.001] | PopD; 0 [PopD; 0] | – | Pop9; 0.018 | PopB; 0.001 |
| PopB | 30 | 5 | 0.931 | 1 | – | 0.33 | 25 | 0.08 | Pop8; 0 | PopA; 0.005 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | 0.971 | 1 | 0.02 | 0.39 | 8 | 0.2 | Pop8; 0.18 [PopF; 0.066] | Pop5; 0.005 [PopD; 0.005] | 60 | PopF; 0.066 | PopD; 0.005 |
| PopD | 21 | 3 | 0.989 | 1 | 0.57 | 0.46 | 18 | 0.08 | Pop4; 0 | Pop6; 0.005 | – | – | – |
| PopE† | 60 | 30 | 0.966 | 1 | 0.98 | 1 | 61 | 0.35 | PopD; 0 | PopF; 0.008 | – | – | – |
| PopF† | 100 | 25 | 0.714 | 1 | 0.92 | 0.3 | 76 | 0.38 | Pop8; 0.18 | Pop2; 0.01 | – | – | – |
| **20 HAPLOTYPES** | | | | | | | | | | | | | |
| PopA | 36 | 45 | 0.982 | 1 | 0.98 | 0.07 | 35 | 0.42 | PopB; 0.001 | PopD; 0 | – | – | – |
| PopB | 30 | 5 | 0.96 | 1 | – | 0.33 | 28 | 0.07 | Pop8; 0 | PopA; 0.005 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | 0.975 | 1 | 0.02 | 0.38 | 8 | 0.21 | Pop8; 0.18 [PopF; 0.066] | Pop5; 0.005 [PopD; 0.005] | 61 | PopF; 0.066 | PopD; 0.005 |
| PopD | 21 | 3 | 0.992 | 1 | 1 | 0.46 | 19 | 0.08 | Pop4; 0 | Pop6; 0.005 | – | – | – |
| PopE† | 60 | 30 | 0.975 | 1 | 0.19 | 0.99 | 66 | 0.36 | PopD; 0 | PopF; 0.008 | – | – | – |
| PopF† | 100 | 25 | 0.812 | 1 | 0.77 | 0.2 | 97 | 0.38 | Pop4; 0.048 | Pop2; 0.01 | – | – | – |

**Table S6:** Inferred dates, proportions of admixture and source groups (as well as the $F_{ST}$ between inferred source groups and the true source), inferred using only 10 or 20 haplotypes from each listed population, in the "hard analysis" scenario. All columns are labeled as in Table S5. In the first column, $^2$ and $^M$ denote populations for which our model's inferred conclusion is "one date, multiway" and "multiple-dates", respectively. †Dates/proportions/sources for these populations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these simulations were inferred using the "multiple-date" grid; see Note S4.3.1. See Table S3 for details of the admixed populations PopA-PopF.

### S5.5.2 robustness to phasing

For both the "full" and "hard" analyses, we tested the effect of phasing on our inference by treating the haplotypes output by the simulator as if phase was unknown. Separately for each chromosome, we combined all 850 haplotypes (50 haplotypes from each of the 17 populations) into a single dataset and used SHAPEIT (52; 53) to generate new phased haplotypes, using default values for all MCMC settings and input parameters. To also test the effectiveness of our method to the wrong genetic map, we input the genetic map from deCode (54) into SHAPEIT for phasing, despite having used the HapMap Phase 2 build 36 genetic map to simulate our data. This latter deCode map was also used when performing inference using CHROMOPAINTER and GLOBETROTTER. For this analysis, we used a slightly reduced set of SNPs (440,400) that were selected to cover the same physical positions as the deCode map.

The results are provided in Table S7. Performance is very similar to the results of Table S5, in terms of dating, proportion and source accuracy. Indeed it performs better in some instances, for example identifying multiway admixture in PopA in both the "full" and "hard" analyses, though the second event was missed in the original "hard" analysis (see Table S5). This suggests that the inferred phase from currently available software is generally an accurate enough representation of the true phase for these settings.

| | | | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation** | **true date** | **true %** | $p$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** | **date** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** |
| **FULL ANALYSIS** | | | | | | | | | | | | | | |
| Pop1 | – | – | 0.03 | – | – | – | – | – | – | – | – | – | – | – |
| Pop2 | – | – | 0.28 | – | – | – | – | – | – | – | – | – | – | – |
| Pop3 | – | – | 0.14 | – | – | – | – | – | – | – | – | – | – | – |
| Pop4 | – | – | 0.14 | – | – | – | – | – | – | – | – | – | – | – |
| Pop5 | – | – | 1 | – | – | – | – | – | – | – | – | – | – | – |
| Pop6 | – | – | 0.23 | – | – | – | – | – | – | – | – | – | – | – |
| Pop7 | – | – | 1 | – | – | – | – | – | – | – | – | – | – | – |
| Pop8 | – | – | 0.85 | – | – | – | – | – | – | – | – | – | – | – |
| Pop9 | – | – | 0.66 | – | – | – | – | – | – | – | – | – | – | – |
| Pop10 | – | – | 0.26 | – | – | – | – | – | – | – | – | – | – | – |
| Pop11 | – | – | 0.13 | – | – | – | – | – | – | – | – | – | – | – |
| PopA$^M$ | 36 | 45 | <.01 | 0.986 | 1 | 0.98 | <.01 | 36 (32-39) | 34 | PopB; 0.001 [Pop8; 0.098] | Pop5; 0 [Pop5; 0] | – | Pop8; 0 | Pop7; 0 |
| PopB$^1$ | 30 | 5 | <.01 | 0.979 | 1 | – | 0.65 | 28 (24-33) | 8 | Pop8; 0 | Pop7; 0 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | <.01 | 0.983 | 1 | 0.02 | 0.69 | 9 (6-11) | 16 | Pop3; 0 [Pop3; 0] | Pop6; 0 [Pop6; 0] | 69 (50-88) | Pop3; 0 | Pop6; 0 |
| PopD$^1$ | 21 | 3 | <.01 | 0.994 | 1 | – | 0.73 | 20 (16-24) | 4 | Pop4; 0 | Pop5; 0 | – | – | – |
| PopE$^1$† | 60 | 30 | <.01 | 0.993 | 1 | 0.98 | 0.94 | 62 (56-68) | 33 | Pop5; 0 | Pop1; 0 | – | – | – |
| PopF$^1$† | 100 | 25 | <.01 | 0.982 | 1 | 0.79 | 0.82 | 93 (84-107) | 24 | Pop3; 0 | Pop1; 0 | – | – | – |
| **HARD ANALYSIS** | | | | | | | | | | | | | | |
| Pop2 | – | – | 0.34 | – | – | – | – | – | – | – | – | – | – | – |
| Pop4 | – | – | 0.1 | – | – | – | – | – | – | – | – | – | – | – |
| Pop9 | – | – | 0.28 | – | – | – | – | – | – | – | – | – | – | – |
| Pop10 | – | – | 0.31 | – | – | – | – | – | – | – | – | – | – | – |
| Pop11 | – | – | 0.11 | – | – | – | – | – | – | – | – | – | – | – |
| PopA$^M$ | 36 | 45 | <.01 | 0.985 | 1 | 0.99 | <.01 | 34 (31-37) | 36 | PopB; 0.001 [PopB; 0.001] | PopD; 0 [PopD; 0] | – | Pop10; 0.017 | PopB; 0.001 |
| PopB$^1$ | 30 | 5 | <.01 | 0.965 | 1 | – | 0.35 | 26 (22-30) | 7 | Pop8; 0 | PopA; 0.005 | – | – | – |
| PopC$^2$‡ | 8 / 58 | 5 | <.01 | 0.982 | 1 | 0.01 | 0.35 | 8 (4-11) | 13 | Pop8; 0.18 [Pop4; 0.048] | Pop6; 0 [PopD; 0.005] | 50 (32-72) | Pop4; 0.048 | PopD; 0.005 |
| PopD$^1$ | 21 | 3 | <.01 | 0.993 | 1 | 0.5 | 0.36 | 21 (18-25) | 7 | Pop4; 0 | Pop6; 0.005 | – | – | – |
| PopE$^1$† | 60 | 30 | <.01 | 0.995 | 1 | 1 | 0.96 | 58 (54-63) | 35 | PopD; 0 | Pop1; 0 | – | – | – |
| PopF$^1$† | 100 | 25 | <.01 | 0.957 | 1 | 0.98 | 0.31 | 92 (78-108) | 17 | Pop3; 0 | Pop2; 0.01 | – | – | – |

**Table S7:** Inferred dates, proportions of admixture and source groups (as well as the $F_{ST}$ between inferred source groups and the true source), inferred after phasing all analysed samples jointly, in both the "full analysis" and "hard analysis" scenarios. All columns are labeled as in Table S5. Our model's inferred conclusion for each population is given next to its name in the first column, with $^1$= "one-date", $^M$= "one date, multiway", and $^2$= "multiple-dates". †Dates/proportions/sources for these populations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these simulations were inferred using the "multiple-date" grid; see Note S4.3.1. See Table S3 for details of the admixed populations PopA-PopF.

### S5.5.3 effect of population bottlenecks following admixture

We used a forwards-in-time simulator to test the effect of severe bottlenecks on our inference, considering a variety of haploid population sizes $N_{\text{pop}}$ and admixture dates $\lambda$.

For each combination of $N_{\text{pop}}$ and $\lambda$ (details below), we simulated an admixture event involving only a small number of admixing individuals by randomly sampling 250 haploid genomes (i.e. 125 individuals) with replacement from a pool of 100 haploids from Pop8 and 150 haploids from Pop2, to yield an overall admixture fraction of 40%, and corresponding to a population bottleneck at the time of admixture. Then for each haploid of the next generation, i.e. the first generation after admixture, we randomly sampled two distinct (parent) haploids from this pool, composing the new haploid's genome as a mosaic of chunks from these two parent haploids, with switches in the mosaic based on the HapMap Phase 2 genetic map as described below. We sampled three different numbers of haploids $N_{\text{pop}}$, denoted "exp", "const", and "shrink" below, by randomly sampling with replacement from this pool of 250 haploids in the first generation, requiring that each new haploid have two distinct parents from the previous generation. For all subsequent $\lambda - 1$ generations, $N_{\text{pop}}$ new haploids were each composed of a mosaic of chunks from two distinct haploids randomly sampled with replacement from the previous generation. After $\lambda$ generations, we randomly sampled 50 haplotypes to form 25 individuals for subsequent analysis.

To simulate recombinations at each generation on each chromosome of each new haploid, we first determined the number of recombination breakpoints $B \equiv B_1 + B_2$. Here $B_1$ is a random sample from a Bernoulli distribution with probability 0.5, and $B_2$ is a random sample from a Poisson distribution with rate equal to the total genetic length of the chromosome in Morgans minus 0.5. $B_1$ models the expected obligate crossover per generation on a chromosome, and $B_2$ models the remaining crossovers. This may represent a more realistic model of mammalian recombination, and certainly differs from the assumption of a simple Poisson model of breakpoints as in our model of ancestry segments. We then sampled the physical location of each of the $B$ breakpoints independently according to their relative genetic map value. Segments on either side of a breakpoint are copied without mutation from the haploid's two different parents.

We simulated three different dates of admixture $\lambda$ as noted in Table S8.

| Simulation | Scenario | Date ($\lambda$) | Kin ("exp" / "const" / "shrink") |
|---|---|---|---|
| **PopG** | 60% Pop2 + 40% Pop8 | 45 | 0.056 / 0.096 / 0.177 |
| **PopH** | 60% Pop2 + 40% Pop8 | 20 | 0.05 / 0.086 / 0.125 |
| **PopI** | 60% Pop2 + 40% Pop8 | 10 | 0.047 / 0.066 / 0.088 |

**Table S8:** Details of admixed populations generated using a forward simulator that mixed Pop2 and Pop8 (see Figure S11) from the "coalescent-based" simulations. Dates are given in generations from present. "Scenario" provides the proportions of admixture from each of Pop2 and Pop8 used to simulate admixture. "Kin" gives a "kinship" estimate inferred using PLINK v1.07 (55) (http://pngu.mgh.harvard.edu/purcell/plink/) by calculating the mean of all pairwise PI_HAT statistics across individuals within a population (see http://pngu.mgh.harvard.edu/~purcell/plink/ibdibs.shtml for more details of PI_HAT as defined in PLINK), providing a measure of the average relatedness of individuals within the population.

Each of Pops G-I were simulated using three different $N_{\text{pop}}$, representing an instantaneous expansion ("exp") to 1250 individuals, i.e. the size of a small town (which we would expect to most closely match the assumptions of our approach of no coalescence since admixture), a constant population size of 125 individuals ("const") corresponding to perhaps a small village, and a very small population ("shrink") of only 50 individuals, immediately following the admixture event. The corresponding haploid population sizes then remain constant until the present day:

1. "exp" ($N_{\text{pop}} = 2500$)
2. "const" ($N_{\text{pop}} = 250$)
3. "shrink" ($N_{\text{pop}} = 100$)

As each $N_{\text{pop}}$ was simulated with one of 3 different admixture dates $\lambda$, this gave 9 simulated populations in total. We note that "const" and "shrink" correspond to population sizes much smaller than likely for most human groups. To evaluate the effect of these bottlenecks on genetic diversity within a population, we calculated the average relatedness of individuals within each of these 9 simulated groups using a pairwise Identity By Descent (IBD) analysis in PLINK v1.07 (55) (http://pngu.mgh.harvard.edu/purcell/plink/). For comparison, we also calculated these values for each of our 95 sampled real data populations. Specifically, to remove any potential effects of linkage disequilibrium, we first thinned SNPs across all populations using a 500kb sliding window, such that the genotypes of any SNP pair within 500kb of each other had squared correlation coefficient (i.e. $r^2$) less than 0.2. This left 73,059 SNPs total across populations, which were then used in PLINK to infer a within population "kinship" estimate by calculating the mean of all pairwise PI_HAT statistics across individuals within a population (see http://pngu.mgh.harvard.edu/∼purcell/plink/ibdibs.shtml for more details of PI_HAT as defined in PLINK). This gave the values in Table S8 for each of our 9 simulated groups.

In these simulations, the kinship estimates range from 0.177 down to 0.047, greater than those for all (100%) and 74%, respectively, of our 80 real-life populations with at least one analysis concluding some evidence of admixture. So these simulations represent bottleneck events more severe than typically, sometimes ever, seen in the real data. Moreover, as we see below our model only performs unreliably for the two strongest bottlenecks, and only for admixture 45 generations (about 1300 years) old, of the 9 simulated cases. This may not be surprising, because in the most extreme case around 36% (16.5% for the second strongest case) of pairs of lineages coalesce more recently than the admixture event, preventing recombination since admixture from fully breaking down LD, even at very long ranges in the genome. Only three of these 80 real-data populations (3.8%) in whom we infer and attempt to characterize admixture: Pima, Colombian and Melanesian have kinship estimates higher than our second strongest simulated bottleneck. Two of these cases involve recent admixture, which we find to be more robustly inferred even in the presence of admixture.

In addition to sharing large numbers of exact genetic segments with each other, these simulations had an additional complication of potentially sharing exact inherited segments with admixed populations PopA and PopB, which were used as donors in our analysis. Specifically, 50 of the 100 haplotypes used to simulate Pops G-I overlap with the 150 haplotypes from Pop8 that were used to simulate Pops A-B.

For each combination of Pops G-I and $N_{\text{pop}}$, we tested separately for admixture using populations from the "hard analysis" as donors. Analogous to our simulations in Note S5.1, to avoid a computational burden, we did not allow any of the donor populations to copy from any of Pops G-I, nor did we allow for "self-copying" in Pops G-I. Otherwise our analyses follows the protocol outlined in Notes S4.1-S4.6.

The results of our analysis are given in Table S9, with the first 9 rows of that table giving results under our usual approach. For each of the 9 simulations, we correctly infer the admixture as "one-event" and precisely infer the admixing groups. Our inferred admixture proportions range from 35-46%, close to the true simulated value of 40%. However, our date estimates for two cases, PopG-const and PopG-shrink, are inferred to be 31-37 and 22-27 generations, respectively, where even the respective upper bounds are modestly (18%) or seriously (40%) below the true value of 45 generations. We also see a very slight underestimation in the dating of PopH-const and PopH-shrink, inferring a date of 15-19 generations for each though the truth is 20 generations.

Theoretical considerations suggest that strong bottlenecks immediately following an admixture event, by generating very long-range LD, can have the effect of making the admixture appear more recent under our approach, with the effect increasing the longer the time since admixture. Note that this is the trend we see here, as estimation is very reliable in PopI for admixture simulated 10 generations ago and very close to the truth in PopH for admixture simulated 20 generations ago, regardless of the severity of the bottleneck. Only populations with very severe bottlenecks, i.e. such that the effective population size is 125 or fewer individuals, will have unreliable date estimation under our standard approach, and then only if the admixture is older than ≈20 generations.

When dividing our coancestry curves by a "NULL" individual constructed to remove any putative admixture signal, exactly as described in Note S4.7 and used to calculate our $p$-values for testing for *any* evidence of admixture, we note that the date inference is extremely accurate in all 9 populations (bottom 9 rows of Table S9). Again, theoretical considerations predict this to be the case, as this procedure automatically accounts for long-range linkage disequilibrium patterns in our coancestry curves that are attributable to the bottleneck, because these are shared across individuals. However, we observe that this robustness in dating comes at the cost of shrinking the heights (or intercepts) of our coancestry curves, which makes inferring source groups more challenging, while our normalised curves more generally have additional noise introduced by the normalisation procedure itself. This is reflected in our slightly less accurate admixture proportion estimation when standardizing our coancestry curves by the "NULL" individual. For example, for the two most difficult scenarios PopG-reg and PopG-shrink, our proportion estimates are 33% and 32%, respectively, after standardizing, compared to 35% and 37%, respectively, before standardizing.

Thus neither form of the coancestry curve leads to completely satisfactory results in these strong-bottleneck cases, but differing estimated admixture dates at least offer a means to diagnose this issue. Therefore, as stated in Note S4.6 we categorize any admixture detected in our real data populations as "uncertain" if the 95% confidence intervals for the dates do not overlap when performing analysis before and after standardizing by the "NULL" individual as described in Note S4.7 (and note that the "NULL" estimated admixture date range is likely to be more reliable, though broader, in such cases). This operates naturally alongside our similarly checking consistency between the curve generation methods to test evidence of admixture, evidence of two admixture dates, and consistency of results with at most two groups admixing (Note S4).

This criterion would correctly flag up PopG-reg and PopG-shrink as having unreliable inference among these 9 simulations, while keeping all others. (For Pops A-F, the other simulated admixed populations in this section, conclusions would not change from those presented in Tables S5 and S7.) Of our 80 real data populations where we would otherwise conclude a signal of characterizable admixture, this criteria affected only three: Biaka Pygmy, Mbuti Pygmy, and Japanese. We describe the effect of this phenomenon in other real data populations in Note S6.4.5.

| | | | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Simulation** | **true date** | **true %** | $p$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** | **date** | **source 1; $F_{ST}$** | **source 2; $F_{ST}$** |
| **NOT STANDARDIZING BY "NULL" INDIVIDUAL** | | | | | | | | | | | | | | |
| PopG-exp[1] | 45 | 40 | <.01 | 0.995 | 1 | – | 0.8 | 40 (37-43) | 39 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopG-const[1] | 45 | 40 | <.01 | 0.993 | 1 | – | 0.95 | 34 (31-37) | 35 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopG-shrink[1] | 45 | 40 | <.01 | 0.959 | 0.999 | – | 0.95 | 25 (22-27) | 37 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-exp[1] | 20 | 40 | <.01 | 0.999 | 1 | 1 | 0.93 | 20 (19-22) | 39 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-const[1] | 20 | 40 | <.01 | 0.998 | 1 | 0.87 | 0.86 | 17 (15-19) | 45 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-shrink[1] | 20 | 40 | <.01 | 0.998 | 1 | 0.61 | 0.97 | 17 (15-19) | 35 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-exp[1] | 10 | 40 | <.01 | 0.998 | 1 | 1 | 0.94 | 10 (9-12) | 41 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-const[1] | 10 | 40 | <.01 | 0.999 | 1 | 1 | 0.84 | 10 (9-11) | 46 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-shrink[1] | 10 | 40 | <.01 | 1 | 1 | 1 | 0.97 | 10 (8-11) | 36 | Pop8; 0 | Pop2; 0 | – | – | – |
| **STANDARDIZING BY "NULL" INDIVIDUAL** | | | | | | | | | | | | | | |
| PopG-exp | 45 | 40 | – | 0.996 | – | – | – | 42 (39-45) | 39 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopG-const | 45 | 40 | – | 0.995 | – | – | – | 45 (41-49) | 33 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopG-shrink | 45 | 40 | – | 0.993 | – | – | – | 44 (40-50) | 32 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-exp | 20 | 40 | – | 0.999 | – | – | – | 21 (18-22) | 39 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-const | 20 | 40 | – | 0.998 | – | – | – | 18 (16-21) | 45 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopH-shrink | 20 | 40 | – | 0.999 | – | – | – | 20 (17-22) | 35 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-exp | 10 | 40 | – | 0.998 | – | – | – | 10 (8-11) | 41 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-const | 10 | 40 | – | 0.999 | – | – | – | 10 (8-11) | 46 | Pop8; 0 | Pop2; 0 | – | – | – |
| PopI-shrink | 10 | 40 | – | 0.999 | – | – | – | 10 (9-12) | 36 | Pop8; 0 | Pop2; 0 | – | – | – |

**Table S9:** Inferred dates, proportions of admixture and source groups (as well as the $F_{ST}$ between inferred source groups and the true source) for nine populations simulated with severe bottlenecks, with inference performed under the "hard analysis" scenario. All columns are labeled as in Table S5. The top nine rows give results without standardizing by the "NULL" individual (i.e. the usual analyses we report); the bottom nine rows gives results after standardization. For the top nine rows, our model's inferred conclusion for each population is given next to its name in the first column, which is [1]= "one-date" (which is the truth here). See Note S5.5.3 for more details of the different bottleneck scenarios and these results.

# S6 Analysis of sample collection

## S6.1 Details of the dataset and phasing

We analysed the autosomal regions of world-wide population sample collections described in (22), (24) and (23), plus 202 samples of anonymous donors drawn from 17 different locations in previously analysed sample collections (21) which we genotyped using the Illumina 660W array. Individuals and SNPs with call rates $< 98\%$ were removed. In total we analyzed 474,491 autosomal SNPs in each of 1530 individuals sampled across 95 different labeled populations, with sample sizes ranging from 2 to 46 individuals per population (prior to removing some outlying samples based on studying the finestructure results – see Note S6.2). See Table S10 and Figure S12 for details of all populations included in this study. The terms used to refer to the populations investigated in this work are as geographically, historically and/or linguistically correct as possible and are chosen with no intention to offend any population group. Generally, to avoid confusion and to maintain consistency across studies, we have kept the original nomenclature of populations from the original papers that published the data (22; 24; 23). In particular, the use of the word "Bantu" throughout the main text and SOM is used for simplicity and always within a linguistic context: BantuSA stands for Bantu-speaking South Africans, and BantuKenya describes Bantu speakers from Kenya. Likewise, our use of "Bantu" as a descriptor for one of the major clades of the fineSTRUCTURE analysis refers to Bantu speakers. Similarly, our use of "Arab" denotes populations that consist mainly of Arabic speakers, or that have historically been part of the Arabian peninsular. We also use the term "Slavic" as a general descriptor for Eastern European populations containing Slavic speakers (56).

As the algorithm described below requires haplotype information, we first phased all populations together running IMPUTEv2 (25) with $Ne = 15000$ and otherwise default settings ($k = 80$, $k\_hap = 150$, 20 posterior MCMC samples after 10 burn-in iterations), using all inferred build 36 haplotypes from Phase 3 of the HapMap project as a reference panel and the build 36 genetic map from Phase 2 of the HapMap project as the fixed recombination rate (reference haplotypes and genetic map available at http://hapmap.ncbi.nlm.nih.gov/ and on the IMPUTEv2 webpage). We used the best-guess (consensus) haplotypes for all further analyses, which also contained imputed values for any missing genotypes. Additional samples of 4 Ashkenazi Jewish (23) and 1 Latino individual, as well as previously unpublished collections of 19 individuals from Croatia and 20 individuals sampled from Dagestan (to be analyzed elsewhere) were phased jointly with the samples depicted in Table S10.

For reasons of computational complexity, chromosomes were phased in 7Mb chunks with a 1Mb buffer at each end. For each individual, the phased haplotypes between each pair of contiguous 7Mb regions were then stitched together by taking the haplotype pairing with the highest number of allele matches in the 1Mb buffer, with an allele's contribution to this match-based score weighted by the inverse of its physical distance from the 1Mb buffer's midpoint.

Subsequently, an initial run of CHROMOPAINTER identified two Ethiopian Jewish samples with poor data quality. Both samples were removed from the analysis.

| **Population Label** | **Geographic Region** | **Source** | **Number of Individuals Analyzed** | **Number of Individuals Removed** |
|---|---|---|---|---|
| Adygei | W.Asia | HGDP | 17 | – |
| Armenian | W.Asia | Behar et al, 2010 | 16 | – |
| Balochi | C.SouthAsia | HGDP | 21 | 3 |
| BantuKenya | Bantu | HGDP | 11 | – |
| BantuSouthAfrica | Bantu | HGDP | 8 | – |

Continued on next page

**Table S10 – continued from previous page**

| Population Label | Geographic Region | Source | Number of Individuals Analyzed | Number of Individuals Removed |
|---|---|---|---|---|
| Basque | N.W.Europe | HGDP | 24 | – |
| Bedouin | S.MiddleEast | HGDP | 45 | – |
| Belorussian | E.Europe | Behar et al, 2010 | 8 | – |
| BiakaPygmy | C.Africa | HGDP | 21 | – |
| Brahui | C.SouthAsia | HGDP | 23 | 2 |
| **Bulgarian** | **E.Europe** | **this study** | **18** | – |
| Burusho | C.SouthAsia | HGDP | 25 | – |
| Cambodian | S.EastAsia | HGDP | 10 | – |
| Chuvash | E.Europe | Behar et al, 2010 | 16 | 1 |
| Colombian | Americas | HGDP | 7 | – |
| Cypriot | W.Asia | Behar et al, 2010 | 12 | – |
| Dai | S.EastAsia | HGDP | 10 | – |
| Daur | N.EastAsia | HGDP | 9 | – |
| Druze | W.Asia | HGDP | 42 | – |
| **EastSicilian** | **S.Europe** | **this study** | **10** | – |
| Egyptian | S.MiddleEast | Behar et al, 2010 | 10 | 2 |
| **English** | **N.W.Europe** | **this study** | **6** | – |
| Ethiopian | Ethiopian | Behar et al, 2010 | 19 | – |
| EthiopianJew | Ethiopian | Behar et al, 2010 | 11 | – |
| **Finnish** | **E.Europe** | **this study** | **2** | – |
| French | N.W.Europe | HGDP | 28 | – |
| Georgian | W.Asia | Behar et al, 2010 | 20 | – |
| **GermanyAustria** | **N.W.Europe** | **this study** | **4** | – |
| **Greek** | **S.Europe** | **this study** | **20** | – |
| Hadza | C.Africa | Henn et al, 2011 | 3 | – |
| Han | S.EastAsia | HGDP | 34 | – |
| HanNchina | S.EastAsia | HGDP | 10 | – |
| Hazara | C.SouthAsia2 | HGDP | 22 | – |
| Hezhen | N.EastAsia | HGDP | 8 | – |
| Hungarian | E.Europe | Behar et al, 2010 | 18 | 2 |
| Indian | C.SouthAsia | Behar et al, 2010 | 13 | 3 |
| IndianJew | C.SouthAsia | Behar et al, 2010 | 8 | – |
| Iranian | W.Asia | Behar et al, 2010 | 13 | 7 |
| **Ireland** | **N.W.Europe** | **this study** | **7** | – |
| Japanese | N.EastAsia | HGDP | 28 | – |
| Jordanian | S.MiddleEast | Behar et al, 2010 | 18 | 2 |
| Kalash | C.SouthAsia | HGDP | 23 | – |
| Karitiana | Americas | HGDP | 14 | – |
| Lahu | S.EastAsia | HGDP | 8 | – |
| Lezgin | W.Asia | Behar et al, 2010 | 18 | – |
| Lithuanian | E.Europe | Behar et al, 2010 | 10 | – |
| Makrani | C.SouthAsia | HGDP | 22 | 3 |
| Mandenka | W.Africa | HGDP | 22 | – |
| Maya | Americas | HGDP | 21 | – |
| MbutiPygmy | C.Africa | HGDP | 13 | – |
| Melanesian | Oceania | HGDP | 10 | – |
| Miao | S.EastAsia | HGDP | 10 | – |
| Mongola | N.EastAsia | HGDP | 10 | – |
| **Moroccan** | **N.Africa** | **Behar et al, 2010; this study** | **22** | **3** |
| Mozabite | N.Africa | HGDP | 25 | 4 |
| Myanmar* | S.EastAsia | Behar et al, 2010 | 3 | – |
| Naxi | S.EastAsia | HGDP | 8 | – |
| NorthItalian | S.Europe | HGDP | 12 | – |
| **Norwegian** | **N.W.Europe** | **this study** | **18** | – |
| Orcadian | N.W.Europe | HGDP | 15 | – |
| Oroqen | N.EastAsia | HGDP | 9 | – |

Continued on next page

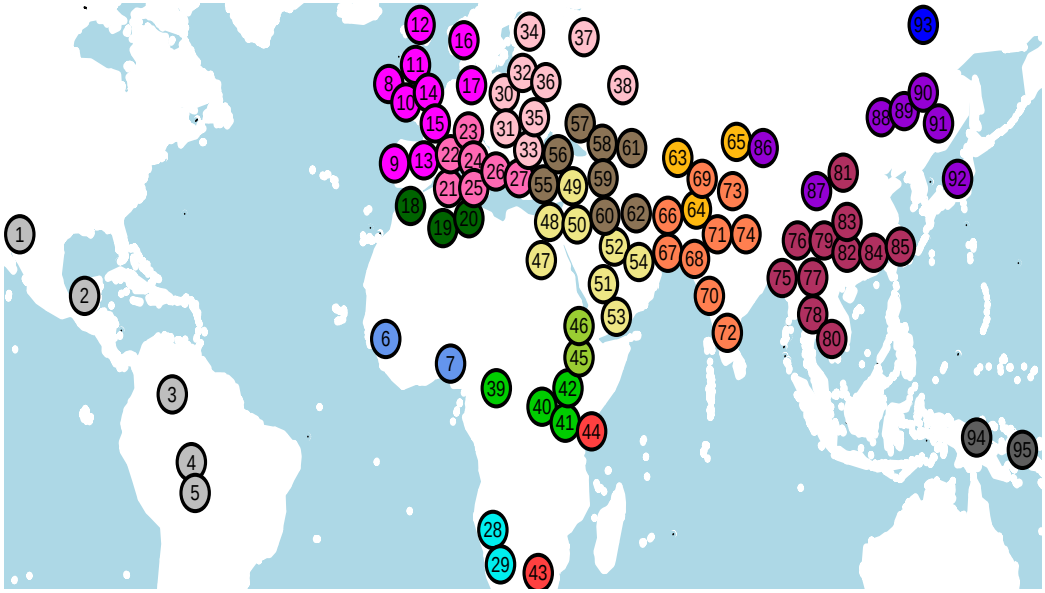
**Table S10 – continued from previous page**

| Population Label | Geographic Region | Source | Number of Individuals Analyzed | Number of Individuals Removed |
|---|---|---|---|---|
| Palestinian | S.MiddleEast | HGDP | 46 | – |
| Papuan | Oceania | HGDP | 16 | 1 |
| Pathan | C.SouthAsia | HGDP | 22 | – |
| Pima | Americas | HGDP | 14 | – |
| **Polish** | **E.Europe** | **this study** | **16** | – |
| Romanian | E.Europe | Behar et al, 2010 | 13 | 1 |
| Russian | E.Europe | HGDP | 25 | – |
| Sandawe | C.Africa | Henn et al, 2011 | 28 | – |
| SanKhomani | San | Henn et al, 2011 | 30 | – |
| SanNamibia | San | HGDP | 5 | – |
| Sardinian | S.Europe | HGDP | 28 | – |
| Saudi | S.MiddleEast | Behar et al, 2010 | 10 | – |
| **Scottish** | **N.W.Europe** | **this study** | **6** | – |
| She | S.EastAsia | HGDP | 10 | – |
| Sindhi | C.SouthAsia | HGDP | 23 | 1 |
| **SouthItalian** | **S.Europe** | **this study** | **18** | – |
| **Spanish** | **N.W.Europe** | **Behar et al, 2010; this study** | **34** | – |
| Surui | Americas | HGDP | 8 | – |
| Syrian | S.MiddleEast | Behar et al, 2010 | 16 | – |
| Tu | N.EastAsia | HGDP | 10 | – |
| Tujia | S.EastAsia | HGDP | 10 | – |
| **Tunisian** | **N.Africa** | **this study** | **12** | – |
| Turkish | W.Asia | Behar et al, 2010 | 17 | – |
| Tuscan | S.Europe | HGDP | 8 | – |
| **UAE** | **S.MiddleEast** | **this study** | **9** | **5** |
| Uygur | C.SouthAsia2 | HGDP | 10 | – |
| Uzbekistani | C.SouthAsia2 | Behar et al, 2010 | 15 | – |
| **Welsh** | **N.W.Europe** | **this study** | **4** | – |
| **WestSicilian** | **S.Europe** | **this study** | **10** | – |
| Xibo | N.EastAsia | HGDP | 9 | – |
| Yakut | Yakut | HGDP | 25 | – |
| Yemeni | S.MiddleEast | Behar et al, 2010 | 4 | 3 |
| Yi | S.EastAsia | HGDP | 10 | – |
| Yoruba | W.Africa | HGDP | 21 | – |

**Table S10:** Details of populations analyzed in the study, collected as part of this study (highlighted in bold) or from "HGDP" (22), "Behar et al, 2010" (23), or "Henn et al, 2011" (24). Empty values in the last column means all samples from the given population were analyzed (see Note S6.2). *Myanmar were a population formed out of the original "Indian" population described in (23); we analyzed these individuals as a separate population and labeled the remaining individuals as "Indian" for our study.

## S6.2 Using fineSTRUCTURE results to remove individuals with differential admixture from majority with same population label

We used the program fineSTRUCTURE (8) to classify individuals into clusters based on similarities among their inferred "copying vectors". For this analysis, we used all individuals from all populations outlined in Table S10 except "Ethiopian Jew" and "Indian Jew", plus additional samples from Croatia and Daghestani as described in Note S6.1. We inferred "copying vectors" using the algorithm described in Note S4.1 and Appendix A. (We note that these "copying vectors" differ from those used elsewhere in this manuscript in that this slightly different set of populations was used.) Taking these copying vectors as input, we used 5,000,000



"AMERICAS"
1. Pima
2. Maya
3. Colombian
4. Karitiana
5. Surui

"W.AFRICA"
6. Mandenka
7. Yoruba

"N.W.EUROPE"
8. Ireland
9. Spanish
10. Welsh
11. Scottish
12. Orcadian
13. Basque
14. English
15. French
16. Norwegian
17. GermanyAustria

"N.AFRICA"
18. Moroccan
19. Mozabite
20. Tunisian

"S.EUROPE"
21. WestSicilian
22. Sardinian
23. NorthItalian
24. Tuscan
25. EastSicilian
26. SouthItalian
27. Greek

"SAN"
28. SanNamibia
29. SanKhomani

"E.EUROPE"
30. Polish
31. Hungarian
32. Lithuanian
33. Bulgarian
34. Finnish
35. Romanian
36. Belorussian
37. Russian
38. Chuvash

"C.AFRICA"
39. BiakaPygmy
40. MbutiPygmy
41. Hadza
42. Sandawe

"BANTU"
43. BantuSouthAfrica
44. BantuKenya

"ETHIOPIAN"
45. Ethiopian
46. EthiopianJewII

"S.MIDDLEEAST"
47. Egyptian
48. Palestinian
49. Syrian
50. Jordanian
51. Saudi
52. Bedouin
53. Yemeni
54. UAE

"W.ASIA"
55. Cypriot
56. Turkish
57. Adygei
58. Georgian
59. Armenian
60. Druze
61. Lezgin
62. Iranian

"C.SOUTHASIA2"
63. Uzbekistani
64. Hazara
65. Uygur

"C.SOUTHASIA"
66. Brahui
67. Makrani
68. Sindhi
69. Kalash
70. IndianJew
71. Balochi
72. Indian
73. Burusho
74. Pathan

"S.EASTASIA"
75. Myanmar
76. Naxi
77. Dai
78. Lahu
79. Yi
80. Cambodian
81. HanNchina
82. Miao
83. Tujia
84. Han
85. She

"N.EASTASIA"
86. Xibo
87. Tu
88. Mongola
89. Daur
90. Oroqen
91. Hezhen
92. Japanese

"YAKUT"
93. Yakut

"OCEANIA"
94. Papuan
95. Melanesian

**Figure S12:** The location of all 95 sampled populations used in the study. Colors denote the major geographical regions ("clades") we assigned each population to based on the results of our fineSTRUCTURE analysis, with our code names for each geographic region provided below the map (see Note S6.2).

Monte-Carlo-Markov-Chain iterations of the fineSTRUCTURE algorithm, discarding the first 1,000,000 iterations as "burn-in" and thinning to only include a single posterior sample for every consecutive 10,000 iterations. We then used default settings in fineSTRUCTURE to find the maximum posterior state after additional hill-climbing moves, which gave 201 final inferred clusters, and to build a tree describing the relationships among these clusters (8). (This algorithm was run twice to check for convergence. Visual inspection suggested the results were highly concordant.) The tree is given in Figure S13.

Based on visual inspection of the inferred tree from fineSTRUCTURE, we divided our 95 labeled populations into the 18 major clades defined in Table S11. These clades are used to help highlight admixture signals in Figures 2-4 of the main text.

| **Clade** | **Population labels** |
|---|---|
| "**San**" | San Khomani, San Namibia |
| "**West Africa**" | Mandenka, Yoruba |
| "**Bantu**" | Bantu Kenya, Bantu South Africa |
| "**Central Africa**" | Hadza, Biaka Pygmy, Mbuti Pygmy, Sandawe |
| "**Ethiopian**" | Ethiopian, Ethiopian Jew |
| "**North Africa**" | Moroccan, Mozabite, Tunisian |
| "**Northwest Europe**" | English, GermanyAustria, Ireland, Norwegian, Orcadian, Scottish, Welsh, Basque, French, Spanish |
| "**East Europe**" | Belorussian, Bulgarian, Chuvash, Finnish, Hungarian, Lithuanian, Polish, Romanian, Russian |
| "**South Europe**" | EastSicilian, Greek, North Italian, Sardinian, South Italian, Tuscan, West Sicilian |
| "**West Asia**" | Adygei, Armenian, Cypriot, Druze, Georgian, Iranian, Lezgin, Turkish |
| "**South Middle East**" | Bedouin, Egyptian, Jordanian, Palestinian, Saudi, Syrian, UAE, Yemeni |
| "**Central South Asia**" | Balochi, Brahui, Burusho, Indian, Indian Jew, Kalash, Makrani, Pathan, Sindhi |
| "**Central South Asia2**" | Hazara, Uygur, Uzbekistani |
| "**Southeast Asia**" | Cambodian, Dai, Han, Han Nchina, Lahu, Miao, Myanmar, Naxi, She, Tujia, Yi |
| "**Northeast Asia**" | Daur, Hezhen, Japanese, Mongola, Oroqen, Tu, Xibo |
| "**Yakut**" | Yakut |
| "**Oceania**" | Melanesian, Papuan |
| "**Americas**" | Colombian, Karitiana, Maya, Pima, Surui |

**Table S11:** Visual inspection of the tree generated under our fineSTRUCTURE analysis (see Figure S13) suggested these 18 world-wide "clades" of groups. The clade's label is given in the first column, and the populations assigned to the clade in the second column. These clades are used in Figures S14-S21 and Figures 2-4 of the main paper.

Furthermore, for each labeled population in which all individuals did not merge with one another in the tree prior to merging with individuals from a different population label, we inspected the copying vectors for obvious discrepancies. An example is individuals with the "Indian" label, whom the fineSTRUCTURE tree separated into five distinct clusters. The copying vectors for these five clusters, averaged across individuals within a cluster, are shown in Figure S14. Four of these five clusters merged into a single group in the tree before merging with any other parts of the tree. The other cluster, however, initially merged with a cluster containing 9 Cambodian individuals, and then with other clusters exclusively containing East Asian populations. Figure S14 illustrates that these three individuals clearly have a distinct genetic pattern from the remaining four Indian clusters. It turns out that these individuals, while initially labeled as "Indian" in our dataset, are actually from Myanmar/Burma (23). For

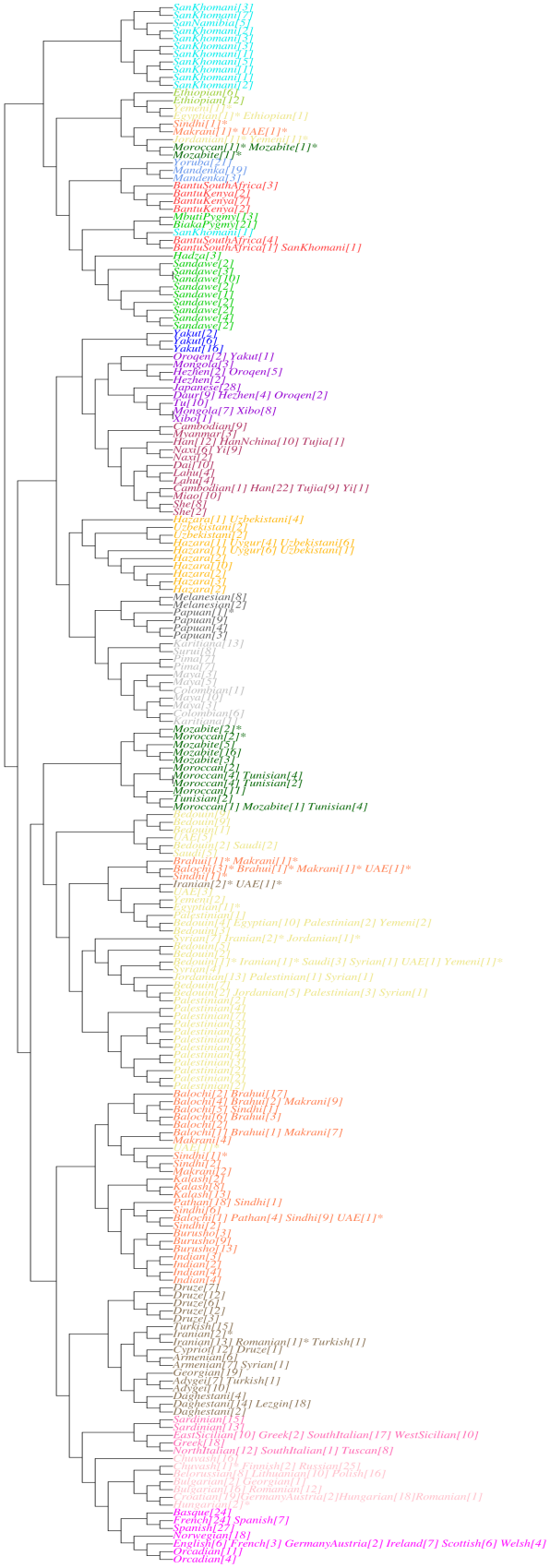



**Figure S13:** FineSTRUCTURE tree results, with leaves colored according to the clade from Table S11 that contains the majority of the leaves' individuals. At each leaf, the number of individuals with the given population label is provided in brackets, with asterisks denoting individuals that were removed from analysis.

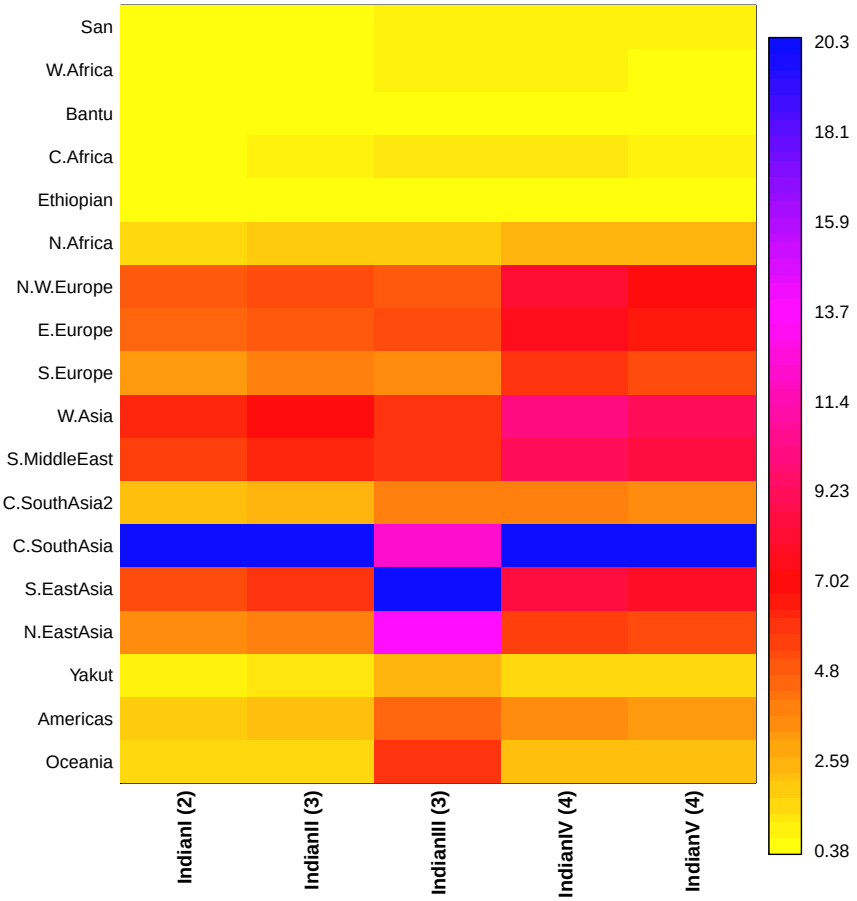



**Figure S14:** The copying vectors for fineSTRUCTURE clusters containing individuals with population label "Indian". Columns give the copying vectors for each cluster averaged across individuals, and the rows give the proportion of DNA copied from each donor group (defined in Table S11), with key given at right. The number of individuals within each cluster is given in parenthesis along the x-axis.

this reason, we divided the "Indian" population into a population of 13 individuals we label "Indian" and a separate population of these 3 individuals that we label "Myanmar", testing each independently for signals of admixture.

For all other populations, we used analogous plots to Figure S14 to remove individuals whose copying vectors appeared to differ visually from the majority of individuals with the same label. The final number of individuals analyzed in each labeled population are given in Table S10, giving 1490 samples in total.

## S6.3 Summary of Results

We will refer to the "full analysis" as our initial (and for many pops only) analysis that allows each of the $K = 95$ recipient populations in Table S10 to copy from every other population in the table. The detailed results from the "full analysis", including the dates and characterization of any inferred admixture events, are provided in the interactive map at http://admixturemap.paintmychromosomes.com/ and summarized in Table S12. See Note S4 for descriptions of all values depicted.

We describe inferred events for some of the populations involved in six "regional analyses" separately in Note S7, and here concentrate on first discussing nine particular groups of events with strong admixture evidence, and then within individual regions some of the populations

with weaker evidence of admixture scores than those mentioned in the main text. We often jointly describe events for populations within the same clade (Table S11). Where appropriate, we offer putative historical explanations for these events, but note that in general these explanations are broad in context and at best provide a guide for interested readers and highlight regions where more focused investigation could follow in the future. In addition, we highlight a small number of cases among the more weakly signalled events, where more detailed inspection of our results suggests additional complexity to that our inference can analyse, or otherwise indicates the results should be treated with caution.

### S6.3.1 summary of nine strongly signaled events

On our webpage at http://admixturemap.paintmychromosomes.com/, we highlight nine sets of admixture events we identify, often showing similar signals across multiple groups. For completeness, we also describe these events here, and note that the descriptions given typically extend those, where present, in the main text. The highlighted cases each include events within the 20 strongest (in terms of coefficient of determination of the top-scoring coancestry curve) of those observed in Table S12, so are likely to be among the events that are the best characterized. However, we note that we are still constrained in our inference of admixing source populations by the available sampled groups. For most events, we discuss historical events that are consistent with our results, in terms of groups involved and importantly also historical date. We note though that these connections are of course not the only possible explanations, given our analysis can look only at current-day populations and given the potential incompleteness of the historical record, particularly in terms of the lasting *genetic* impact of different events, even if other event details are well recorded. However, in some cases (particularly some more recent events, e.g. admixture in the Americas) we do not know of plausible known historical alternative events that might explain the signals seen.

**Colonial-era European and African admixture in the Maya and Pima** Our analysis infers admixture between three groups in the Maya dating to around 1670CE, with distinct inferred admixture sources from Europe (most genetically similar to the Spanish), West Africa (the Yoruba) and the Americas (the Pima, the nearest sampled group in the Americas). These sources are consistent with historically attested colonial era migrants from Spain and West Africa from the 16th Century onwards. The Pima also show evidence of a single admixture event between Maya-like and Eurasian (Turkish)-like sources. The date of this event 1754CE (1698-1810CE) is also consistent with the European colonial-era, although the identity of the Eurasian component is less clearly related to the European colonial nations.

**Admixture in Cambodia dating to the period of the decline and fall of the Indianised Khmer empire** We infer a ≈19% contribution from a source that is related to modern-day Central, South and East Asians, and an ≈81% contribution from a source related specifically to modern-day Han and Dai, the latter a branch of the Tai people who entered the region in historical times (30). This event, with similar proportions to those inferred in (29), is dated to 1362CE (1194CE - 1502CE), a period spanning the end of the Indianized Khmer empire (802-1431AD), one of the most powerful empires in Southeast Asia whose fall was hypothesized to relate to a Tai influx (30).

**Mongol era admixture in Eurasia** The rapid expansion led by Genghis Khan and the subsequent Mongol empire (1206-1368CE; (31)) is one of the most dramatic events in human history. One population believed to be at least partly descended from these Mongols, based on historical, linguistic and oral tradition, and unusual patterns of Y-chromosome male descent, are

the Hazara from Pakistan (32; 33). Using only autosomal genetic data, we independently infer this population to show the clearest signal of admixture in the entire dataset, with an admixture event occurring 22 (19-24) generations in the past, or 1306CE (1250-1390CE), between a source similar to the Iranians (55% contribution), and a source most closely similar among our sampled groups to present-day Mongolians (45%), confirming that both the date and origin of admixture link precisely to the Mongol empire. In fact we find that the Hazara are just one of seven populations (four among the top 20 clearest signals), including the Uyghur (34) and the Mongola themselves, who show an admixture event between a local source and a source closely genetically related to the Mongola, dating within the Mongol Period (Figure 2D, Table S12). These populations were all sampled from within the range of the Mongol expansion and show a progressive westward decrease in Mongol ancestry. We however note that the slightly earlier date in the Turkish of 1250AD (1166-1362) is not inconsistent with other known Turkic pre-Genghis movements from East Asia, such as the Oghuz Turks (31).

The Mongola themselves interestingly show evidence of a more complex admixture history involving admixture between multiple groups at around the same time, with the strongest admixture direction involving primarily Northeast Asian groups on the one side and southern Chinese groups on the other, dated to 1334CE (1194-1446CE). This most prominent event corresponds to expectations based on the history of the Mongol empire (31). A second direction implies admixture involving a third group containing ancestry components more closely related to European populations. Because the Mongols spread right across Asia, this mixture appears highly consistent with the incorporation of DNA from both China and West Eurasia, within their empire, into the Mongolian population during this time, highlighting the potential impact of this period on the genetic ancestry across Eurasia. We note that ostensibly similar admixture signals in terms of groups involved, involving West Eurasian admixture into other groups in N.E. Asia, e.g. the Xibo and the Yakut, in fact date to clearly different times in our inference, and so represent distinct events.

Previously described techniques (27) also identified East Asian admixture in the Uyghur and Hazara using these samples (11), dating it to a similar time period, disagreeing with earlier results that suggested a much older date (34; 57). While these authors also attribute their findings to Mongol Era admixture, they select the Japanese and Italian from among HGDP donors as approximate present-day proxies for the two admixing source groups. In contrast, our method selects the Mongols and Iranians as the best present-day proxies for these admixing sources in both populations, strongly supporting the theory that the major genetic contribution to these populations is linked to the Mongol invasions initiated by Genghis Khan. Our broader collection of populations, and analysis approach, allow us to narrow more subtle admixture signals in five further groups (Turkish, Lezgin, Uzbekistani, Mongola, and Daur) to the date and location of the Mongol Empire.

**Shared admixture events in Eastern Europe in the first Millenium** FineSTRUCTURE did not fully separate groups from Eastern Europe so we repainted them, excluding each other as donors ("East Europe I" analysis; see Note S7.6).

The two most easterly groups, the Russians and Chuvash, show similar signals of inferred admixture between two sources, one with ancestry related to Northeast Asians and one related to Europeans (Table S16). There is evidence in both groups ($p<0.05$) that this occurred at more than one time (Figure 2D), with very ancient East Asian ancestry prior to 500BC, in the case of Russia at least, and a recent event, consistent with approximately Mongol-empire era admixture, together contributing ≈10% of DNA in Russians and ≈35% in the Chuvash.

Six other Eastern European populations show highly shared events (Figure 3). In contrast to Central Asia, none of these populations shows evidence of multiple admixture times ($p>0.5$), but all six independently show evidence for admixture between more than two groups

($p<0.02$). For example in Bulgarian genomes, DNA chunks more closely related to those carried by present-day Greek, Norwegian and Oroqen people tend to be separated in the genome so that each pair of curves shows a dip at short genetic distances (Figure 3), implying these segments must occur on three different ancestry backgrounds. The clearest admixture signal in each population predates the Mongol empire but involves the minority source group having at least some ancestry related to Northeast Asians (e.g. the Oroqen, Mongola and Yakut), with ≈2-4% of these groups' total ancestry proportion linking directly to East Asia, lowest in our Polish sample and highest in Hungarians. A second signal in each group involves admixture between distinct European and West Eurasian groups – one more southerly (sharing more DNA segments with e.g. Greece and W. Asia) and one more northerly (sharing more DNA segments with e.g. North and Northwest Europe) – at approximately the same time as the East Asian admixture, and with each group inferred as contributing relatively similar amounts of DNA (though the fraction is uncertain). Inferred admixture dates fall within a relatively tight range (440-1080CE), earliest in Poland, with groups more distant from Poland generally having more recent dates. Because this signal is subtle, we constructed one of our coalescent-based admixture settings to incorporate simultaneous three-way admixture between similarly differentiated groups, and at a similar time (36 generations), to those inferred here: encouragingly for our method's power, GLOBETROTTER accurately determined all three groups given appropriate source populations (see Notes S5.4-S5.5).

These results are consistent with our detecting a genetic legacy from invasions of peoples from the Asian steppes (e.g. the Huns, Magyar and Turkic Bulgars) during the first millennium CE (31; 38), and affecting all six groups. We speculate that the second event seen in our six Eastern Europe populations between northern European and southern European ancestral sources may correspond to the expansion of Slavic language speaking groups (commonly referred to as the Slavic expansion) across this region at a similar time, perhaps related to displacement caused by the Eurasian steppe invaders (38; 58). Under this scenario, the northerly source in the second event might represent DNA from Slavic-speaking migrants (sampled Slavic-speaking groups are excluded from being donors in the EastEurope I analysis). To test consistency with this, we repainted these populations adding the Polish as a single Slavic-speaking donor group ("East Europe II" analysis; see Note S7.6) and, in doing so, they largely replaced the original North European component (Figure S21), although we note that two nearby populations, Belarus and Lithuania, are equally often inferred as sources in our original analysis (Table S12). Outside these six populations, an admixture event at the same time (910CE, 95% CI:720-1140CE) is seen in the southerly neighboring Greeks, between sources represented by multiple neighboring Mediterranean peoples (63%) and the Polish (37%), suggesting a strong and early impact of the Slavic expansions in Greece, a subject of recent debate (37). These shared signals we find across East European groups could explain a recent observation of an excess of IBD sharing among similar groups, including Greece, that was dated to a wide range between 1,000 and 2,000 years ago (37).

**Sub-Saharan African admixture in populations bordering the Mediterranean Sea, Arabian Sea and the Persian Gulf dating to the period of the Arab slave trade**

Sub-Saharan admixture is seen among 17 populations from the Mediterranean, North Africa (consistent with the findings of (59; 20)) and the Near East, and bordering the Arabian Sea and Persian Gulf, including six of the 20 clearest admixture signals, with five of these 17 populations inferred to show admixture at more than one time (two further groups show signals close to our significance threshold of 0.05). While the inferred donor, or highest-contributing sub-Saharan donor, is West African (Yoruba) for all 12 populations bordering the Mediterranean Sea, Bantu-speaking groups from East or South Africa contribute most strongly for all 5 Persian Gulf populations (Figure 2D). In over 2,400 simulations of admixture involving Yoruba, we observed

100% ability to distinguish this source from the East and South African groups, which together with the consistency of the observed signal suggests this inference is reliable (see Notes S5.1-S5.2). These events influence populations that are either predominantly Arabic speaking, or have a history of Arab trade or occupation. We interpret these signals as resulting from the Arab slave trade, which originated around the 7th century (35), and our event dates are highly consistent with this range and indicate a spread of times even within individual groups, although they also imply earlier sub-Saharan African gene flow into e.g. the Moroccans. Our analyses confirm genetically different sources for the trans-Saharan (supplying the Mediterranean region) and Indian Ocean (supplying the Gulf region) trades (35).

Among these 17 populations, the earliest admixture signal is seen in the Druze, who we infer to be affected by a single admixture event, with just 3% Yoruban-like ancestry dating from 890CE (770-1000CE). This admixture fraction is lower than that seen in the other nearby Middle Eastern groups. The Druze are members of a religious group, founded around 1014CE, which prohibited the keeping of slaves and has been closed to new members since 1043CE (60). Admixture pre-dating formation of this group, and the low admixture fraction seen, are consistent with the early abolition of slavery in the Druze, and subsequent relative genetic isolation of this group from nearby populations.

**Waves of admixture in Southern Africa** The South African Bantu speakers show a genetic contribution from the San (Khomani) dating to 1220CE (1080-1360CE), highlighting genetic influences consistent with the dispersal of Bantu-speaking peoples into southern Africa (61; 62; 63), and placing an upper bound on the arrival time of these peoples in the region consistent with archaeological dates (64; 65).

The strongest evidence of complex admixture in the entire dataset of 95 populations is in the Khomani San from the Northern Cape province of South Africa. Here we infer one older event (of very uncertain age prior to 1000CE) between sources more similar to present-day Namibian San and Bantu speakers. Such an event is complementary to, though potentially slightly earlier than, the signal seen in the South Africa Bantu speakers and demonstrates substantial admixture from Bantu-speaking peoples in this San group. A second, more recent (1700-1840CE) admixture episode occurred between the San Khomani and one group with a mixture of inferred ancestries: North European, and South Asian, contributing an estimated 27% of DNA (Figure 2D). This date is consistent with the colonial-period arrival of European settlers (from Britain, Holland, and Germany) and South Asian immigrants (39). We note we do not (attempt to) directly separate more than two recent sources, although the wide geographic separation between Northwest Europe and Central/South Asia, with only small inferred contributions from groups between these locations is perhaps most simply explained by more than one Eurasian group being involved in the recent event. Thus, we observe a likely genetic impact for two distinct waves of settlement into South Africa, one by Bantu and the other by colonial-era migrants (61).

**Shared ancient and diverse modern admixture in Central Asia** Central Asia shows a particularly complex inferred history (see Figure 4 of main text). Among ten sampled populations, eight are from Pakistan. Several groups – e.g. the Brahui and Balochi – are not well separated by fineSTRUCTURE (Figure S13). Therefore, we repainted each Central Asian population excluding all other Central Asian groups as donors ("Central Asia analysis"; see Note S7.4). Of the 8 groups from Pakistan, three show diverse single admixture, and five show complex admixture in the "Central Asia analysis". The results of this analysis closely match those of the full analysis in the recent events found. Several of the recent events are discussed in other sections (the Makrani, the Brahui and the Sindhi in "Sub-Saharan African admixture and the Arab slave trade"; and the Hazara, who have a signal similar to the Uyghur and Uzbekistani

groups from outside Pakistan, in "Mongol era admixture in Eurasia").

The Kalash are a geographically and genetically (39) isolated population that have lived in a remote valley within present-day Pakistan for many centuries (40; 66). In the original (Full) analysis, the Kalash possess our oldest estimated date of most recent admixture, of 600BCE (990-210BCE), between sources best represented today by Germany-Austria (though within a range of potential European-related sources, e.g. represented by Turkey in the CentralAsia analysis; 35%) and the nearby Pathan (65%). Intriguingly, this period overlaps that of Alexander the Great (356-323BCE) whose army, local tradition holds, the Kalash are descended from (40). The history of this group is not known: our analysis suggests a major admixture event from a source related to present-day Western Eurasians, but we cannot identify the geographic origin of this ancient source precisely.

In the "Central Asia" analysis of Note S7.4 (but not in the "full" analysis), a very similar ancient admixture signal (always dated older than 90BCE) is seen in five nearby Pakistan populations: the Makrani, Balochi, and Brahui, and more weakly in the Pathan and Sindhi, but not identified in the most northerly groups. Ancient admixture involving sources related to East Asia is inferred in the easterly Burusho and tentatively (within a second signal) the Kalash. These older events are similar in date to that seen in the Kalash but involve less strongly European-like, and more West Asian like, sources (Figure 4; Figure S18), and pre-date recorded history for the region.

**Admixture in populations situated near the route of the Silk Road** The Tu from China have admixture (1080-1330CE) inferred to occur between locally related sources, and a source group most closely related to present-day Europeans (e.g. Greeks). We find this European-like component difficult to interpret with certainty, though speculate that it may relate to traders traveling the nearby Silk Road (38). The same might also apply to a very similar but weaker signal (estimated at 6% admixture) seen in the Han from North China.

The presence of Europeans in China before the Middle Age is virtually unknown, while extended contacts initiated and developed in a substantial way during this period. A similar proportion of West Eurasian admixture was identified in the Tu samples using a different approach (27).

**Admixture in North-Eastern China** The Hezhen from northeastern China, who are descended from the Jurchen, show a clear signal of a recent admixture event 1560CE (1470-1640CE) between a group similar to the nearby Oroqen, who speak a related Tungusic language, and a group represented by a mixture of more southerly Chinese populations (49%). This event overlaps the later part of the Ming dynasty, a period of incorporation of the Jurchen into a broader Manchu synthesis of Jurchen, Chinese and Mongol elements (67). The Oroqen themselves show an event (1502CE; 1390-1642CE), between a group most resembling the northerly Yakut from Siberia, and a group like the southerly neighbouring Daur. The Daur have given the Oroqen many loanwords (68) over centuries of contact.

| | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Population** | $n_k$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| **ONE DATE** | | | | | | | | | | | | |
| Hazara$\Delta$ | 22 | 0.996 | 1 | 0.11 | 0.29 | 1306 (1250-1390) | 45 | Mongola | Iranian | – | – | – |
| Uzbekistani$\Delta$ | 15 | 0.992 | 1 | 0.11 | 0.37 | 1390 (1334-1474) | 39 | Mongola | Iranian | – | – | – |
| Uygur$\Delta$ | 10 | 0.989 | 1 | 0.28 | 0.52 | 1306 (1250-1390) | 50 | Mongola | Iranian | – | – | – |
| Makrani$\Delta$ | 22 | 0.988 | 1 | 0.08 | 0.24 | 1418 (1334-1474) | 5 | BantuSA | Balochi | – | – | – |
| Druze | 42 | 0.988 | 0.998 | 0.43 | 0.09 | 886 (774-998) | 3 | Yoruba | Cypriot | – | – | – |
| Mozabite$\Delta$ | 25 | 0.988 | 1 | 0.13 | 0.61 | 1334 (1250-1418) | 8 | Yoruba | Moroccan | – | – | – |
| Turkish | 17 | 0.985 | 0.999 | 0.08 | 0.26 | 1250 (1166-1362) | 8 | Mongola | Iranian | – | – | – |
| Brahui$\Delta$ | 23 | 0.985 | 1 | 0.87 | 0.23 | 1362 (1306-1502) | 2 | BantuKenya | Balochi | – | – | – |
| Yemeni | 4 | 0.982 | 1 | 0.11 | 0.05 | 1530 (1390-1614) | 13 | BantuSA | Syrian | – | – | – |
| Pima | 14 | 0.957 | 0.994 | 0.31 | 0.08 | 1754 (1698-1810) | 12 | Turkish | Maya | – | – | – |
| BantuSA | 8 | 0.955 | 1 | 0.12 | 0.47 | 1222 (1082-1362) | 27 | SanKhomani | Yoruba | – | – | – |
| Tu | 10 | 0.95 | 1 | 0.05 | 0.32 | 1222 (1082-1334) | 9 | Greek | HanNchina | – | – | – |
| WestSicilian$\Delta$ | 10 | 0.941 | 1 | 0.07 | 0.13 | 1166 (914-1362) | 4 | Yoruba | EastSicilian | – | – | – |
| Cambodian | 10 | 0.919 | 0.999 | 0.22 | 0.07 | 1362 (1194-1502) | 19 | Uygur | Han | – | – | – |
| Georgian | 20 | 0.904 | 0.994 | 0.28 | 0.1 | 1082 (914-1278) | 31 | Adygei | Turkish | – | – | – |
| Romanian$\Delta$ | 13 | 0.899 | 1 | 0.82 | 0.36 | 1054 (886-1194) | 43 | Lithuanian | EastSicilian | – | – | – |
| Bulgarian$\Delta$ | 18 | 0.885 | 1 | 0.11 | 0.34 | 1138 (942-1334) | 46 | Belorussian | Cypriot | – | – | – |
| Hezhen | 8 | 0.88 | 0.999 | 0.13 | 0.24 | 1558 (1474-1642) | 49 | Tujia | Oroqen | – | – | – |
| Oroqen | 9 | 0.878 | 0.998 | 0.25 | 0.05 | 1502 (1390-1642) | 23 | Yakut | Daur | – | – | – |
| Hungarian$\Delta$ | 18 | 0.866 | 1 | 0.79 | 0.28 | 830 (662-1026) | 36 | Cypriot | Belorussian | – | – | – |
| HanNchina | 10 | 0.863 | 1 | 0.44 | 0.13 | 1194 (998-1418) | 6 | Turkish | Tujia | – | – | – |
| Daur | 9 | 0.853 | 0.999 | 0.6 | 0.23 | 1334 (1250-1446) | 41 | Tujia | Oroqen | – | – | – |
| Greek$\Delta$ | 20 | 0.823 | 1 | 0.8 | 0.34 | 914 (718-1138) | 37 | Polish | Cypriot | – | – | – |
| Melanesian | 10 | 0.806 | 1 | 0.18 | 0.24 | 1138 (718-1502) | 49 | Papuan | Myanmar | – | – | – |
| Mandenka | 22 | 0.698 | 1 | 0.28 | 0.26 | 1390 (1166-1586) | 29 | Yoruba | Yoruba | – | – | – |
| Indian | 13 | 0.665 | 0.997 | 0.74 | 0.09 | 438 (150B-830) | 14 | Myanmar | Sindhi | – | – | – |
| NorthItalian | 12 | 0.621 | 0.997 | 0.92 | 0.11 | 66B (766B-550) | 33 | Cypriot | Welsh | – | – | – |
| Polish$\Delta$ | 16 | 0.581 | 0.995 | 0.94 | 0.06 | 1054 (718-1278) | 34 | French | Lithuanian | – | – | – |
| Tuscan | 8 | 0.569 | 0.996 | 0.94 | 0.13 | 942 (522-1222) | 41 | Cypriot | French | – | – | – |
| NorthItalian† | 12 | 0.462 | 0.997 | 0.73 | 0.17 | 542B (1886B-606) | 27 | Jordanian | French | – | – | – |
| SanNamibia$\Delta$ | 5 | 0.318 | 1 | 0.99 | 0.24 | 578 (46-1110) | 2 | Sandawe | SanKhomani | – | – | – |
| **ONE DATE, MULTIWAY** | | | | | | | | | | | | |
| Lezgin | 18 | 0.946 | 0.995 | 0.26 | <.01 | 1306 (1138-1446) | 10 | Uygur [IndianJew] | Turkish [Scottish] | – | Oroqen | Georgian |
| IndianJew | 8 | 0.914 | 0.995 | 0.08 | <.01 | 1362 (1166-1530) | 46 | Iranian [Saudi] | Sindhi [Indian] | – | Myanmar | Sindhi |
| Mongola | 10 | 0.931 | 0.996 | 0.29 | <.01 | 1334 (1194-1446) | 36 | Oroqen [Yakut] | HanNchina [Miao] | – | Yemeni | Oroqen |
| Han | 34 | 0.578 | 0.987 | 0.31 | <.01 | 1194 (606-1530) | 20 | Dai [Dai] | HanNchina [HanNchina] | – | She | Cambodian |
| Jordanian$\Delta$ | 18 | 0.992 | 0.997 | 0.26 | <.01 | 1222 (1138-1334) | 6 | Yoruba [Yoruba] | Syrian [Saudi] | – | Palestinian | Georgian |
| Tunisian$\Delta$ | 12 | 0.981 | 0.997 | 0.05 | <.01 | 1334 (1222-1446) | 8 | Yoruba [Yoruba] | Moroccan [Moroccan] | – | Saudi | Mozabite |
| SouthItalian$\Delta$ | 18 | 0.904 | 0.996 | 0.12 | <.01 | 886 (550-1362) | 10 | Egyptian [Mandenka] | WestSicilian [Welsh] | – | English | Yemeni |
| Bedouin$\Delta$ | 45 | 0.993 | 1 | 0.05 | <.01 | 1138 (1082-1222) | 5 | Yoruba [Yoruba] | Jordanian [Saudi] | – | Welsh | Saudi |
| Cypriot | 12 | 0.918 | 0.995 | 0.55 | <.01 | 662 (270-998) | 9 | Egyptian [Hadza] | EastSicilian [Greek] | – | Welsh | Saudi |
| Naxi | 8 | 0.585 | 0.988 | 0.17 | <.01 | 1446 (970-1754) | 39 | Han [Myanmar] | Yi [Yi] | – | Dai | Yi |
| Balochi$\Delta$ | 21 | 0.958 | 0.991 | 0.13 | 0.01 | 1390 (1306-1558) | 43 | Sindhi [Lithuanian] | Brahui [Brahui] | – | Makrani | Brahui |
| BantuKenya | 11 | 0.906 | 0.994 | 0.49 | 0.01 | 1222 (1054-1362) | 30 | Ethiopian [EthiopianJew] | Yoruba [Yoruba] | – | Sandawe | Yoruba |
| Norwegian | 18 | 0.614 | 0.986 | 0.93 | 0.01 | 746 (354-1166) | 8 | Russian [Yakut] | Scottish [Scottish] | – | Lithuanian | Ireland |
| Saudi | 10 | 0.924 | 0.989 | 0.42 | 0.01 | 1278 (1138-1418) | 47 | Bedouin [Bedouin] | Syrian [Syrian] | – | Hadza | Bedouin |

Continued on next page

| Table S12 – continued from previous page | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | first event | | | | second event signal | | |
| **Population** | $n_k$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| IranianΔ | 13 | 0.96 | 0.991 | 0.08 | 0.01 | 1306 (1138-1474) | 2 | BantuKenya [BantuKenya] | Turkish [Georgian] | – | Indian | Saudi |
| SardinianΔ | 28 | 0.767 | 0.998 | 0.75 | 0.01 | 634 (326-830) | 24 | Egyptian [Hadza] | French [Welsh] | – | Basque | SouthItalian |
| Maya | 21 | 0.994 | 0.995 | 0.28 | 0.01 | 1670 (1642-1726) | 19 | Spanish [Basque] | Pima [Colombian] | – | Yoruba | Karitiana |
| SyrianΔ | 16 | 0.986 | 0.999 | 0.13 | 0.01 | 1222 (1110-1362) | 5 | Yoruba [Yoruba] | Iranian [Georgian] | – | Saudi | Yoruba |
| KalashΔ | 23 | 0.778 | 0.988 | 0.16 | 0.01 | 598B (990-206B) | 35 | GermanyAustria [Scottish] | Pathan [Indian] | – | Naxi | Burusho |
| Kalash†Δ | 23 | 0.823 | 0.988 | 0.24 | 0.02 | 738B (1326-94B) | 35 | GermanyAustria [Scottish] | Pathan [Burusho] | – | Burusho | Naxi |
| LithuanianΔ | 10 | 0.765 | 0.996 | 0.71 | 0.02 | 914 (522-1194) | 1 | Daur [Colombian] | Belorussian [Belorussian] | – | Russian | Hadza |
| FinnishΔ | 2 | 0.415 | 1 | 0.64 | 0.02 | 802 (242-1334) | 45 | Russian [Oroqen] | Norwegian [Norwegian] | – | Orcadian | Russian |
| Myanmar | 3 | 0.646 | 0.999 | 0.13 | 0.04 | 1306 (1026-1726) | 31 | Papuan [Papuan] | Han [Cambodian] | – | Colombian | Melanesian |
| BelorussianΔ | 8 | 0.832 | 1 | 0.93 | 0.04 | 1082 (942-1250) | 8 | Uygur [Daur] | Lithuanian [Lithuanian] | – | Basque | Lithuanian |
| **MULTIPLE DATES** | | | | | | | | | | | | |
| SanKhomani‡Δ | 30 | 0.991 | 0.999 | 0.01 | <.01 | 1754 (1698-1838) | 27 | Egyptian [Welsh] | BantuSA [SanNamibia] | 1494B (10762B-970) | SanNamibia | BantuSA |
| Yakut‡ | 25 | 0.955 | 1 | 0.02 | 0.32 | 1670 (1530-1838) | 11 | Russian [Russian] | Oroqen [Oroqen] | 102 (1438B-802) | Russian | Oroqen |
| Moroccan‡Δ | 22 | 0.972 | 0.995 | 0.02 | <.01 | 1502 (1418-1670) | 10 | Mozabite [Yoruba] | Egyptian [Basque] | 710B (1914B-326) | Mozabite | Basque |
| Chuvash‡Δ | 16 | 0.967 | 1 | 0.02 | 0.26 | 1502 (1110-1698) | 41 | Uzbekistani [Yakut] | Finnish [Lithuanian] | 1606B (2558B-298) | Yakut | Lithuanian |
| Spanish‡Δ | 34 | 0.964 | 0.999 | 0.02 | <.01 | 1334 (1054-1586) | 16 | French [Hadza] | French [Basque] | 234B (2530B-382) | BantuKenya | Basque |
| Sandawe‡ | 28 | 0.962 | 1 | 0.02 | 0.26 | 1446 (1222-1558) | 37 | Ethiopian [Hadza] | BantuSA [BantuKenya] | 682B (1886-262B) | BantuKenya | Ethiopian |
| Russian‡Δ | 25 | 0.961 | 1 | 0.03 | 0.27 | 1306 (1054-1530) | 7 | Oroqen [Oroqen] | Polish [Lithuanian] | 2054B (5162-710B) | Oroqen | Lithuanian |
| Xibo | 9 | 0.953 | 0.999 | 0.03 | <.01 | 1810 (1698-1894) | 10 | Uzbekistani [Lithuanian] | Mongola [Daur] | 886 (158-1166) | Oroqen | Naxi |
| UAE‡Δ | 9 | 0.977 | 1 | 0.03 | 0.02 | 1754 (1642-1838) | 4 | BantuSA [BantuKenya] | Saudi [Saudi] | 746 (794B-1110) | BantuKenya | Saudi |
| Sindhi‡Δ | 23 | 0.947 | 0.995 | 0.03 | <.01 | 1782 (1670-1866) | 43 | Balochi [Balochi] | Pathan [Pathan] | 346B (3118B-74) | Brahui | Welsh |
| Adygei‡ | 17 | 0.961 | 1 | 0.03 | 0.03 | 1530 (1418-1810) | 9 | Uygur [Oroqen] | Turkish [Georgian] | 990B (3678B-830) | Oroqen | Georgian |
| Palestinian‡Δ | 46 | 0.992 | 0.999 | 0.03 | 0.23 | 1390 (1250-1586) | 3 | Yoruba [EthiopianJew] | Jordanian [Jordanian] | 382 (1774B-774) | Hadza | Jordanian |
| Ethiopian‡Δ | 19 | 0.93 | 0.998 | 0.03 | 0.01 | 1362 (1222-1586) | 47 | Egyptian [Jordanian] | EthiopianJew [BantuKenya] | 2530B (3930-1018B) | Saudi | EthiopianJew |
| EthiopianJew‡Δ | 11 | 0.879 | 0.999 | 0.04 | 0.03 | 1530 (1026-1670) | 12 | Jordanian [Lahu] | Ethiopian [Ethiopian] | 458B (3762-10B) | Saudi | Ethiopian |
| Pathan‡Δ | 22 | 0.936 | 0.988 | 0.04 | <.01 | 1362 (1110-1642) | 23 | Sindhi [Indian] | Iranian [Scottish] | 2530B (10510-1242B) | Scottish | Sindhi |
| Burusho‡Δ | 25 | 0.973 | 0.998 | 0.04 | 0.08 | 1026 (802-1642) | 19 | Uygur [Yi] | Pathan [Pathan] | 5218B (11910B-214) | Yi | Sindhi |
| EastSicilian‡Δ | 10 | 0.911 | 0.999 | 0.05 | 0.1 | 1474 (1054-1698) | 10 | Egyptian [Mandenka] | WestSicilian [NorthItalian] | 654B (6954B-186) | BantuKenya | Welsh |
| Egyptian‡Δ | 10 | 0.987 | 1 | 0.05 | 0.3 | 1586 (1306-1754) | 10 | Yoruba [Yoruba] | Syrian [Saudi] | 914 (346B-1054) | Yoruba | Saudi |
| **UNCERTAIN** | | | | | | | | | | | | |
| BiakaPygmy | 21 | 0.851 | 0.998 | 0.05 | 0.1 | 1362 (1138-1530) | 46 | BantuSA | Yoruba | – | – | – |
| MbutiPygmy | 13 | 0.773 | 0.994 | – | <.01 | 1530 (1418-1642) | 42 | Yoruba | BantuSA | – | – | – |
| Japanese | 28 | 0.701 | 0.993 | 0.19 | <.01 | 1222 (942-1446) | 9 | Uygur | HanNchina | – | – | – |
| Lahu† | 8 | 0.316 | 0.982 | 0.63 | 0.14 | 18 (1550B-1194) | 25 | Naxi | Cambodian | – | – | – |
| Lahu | 8 | 0.302 | 0.982 | 0.58 | 0.13 | 326 (486B-1698) | 49 | Cambodian | Yi | – | – | – |
| Hadza | 3 | 0.738 | 0.978 | 0.46 | <.01 | 1558 (1334-1698) | 24 | BantuSA | Sandawe | – | – | – |
| French | 28 | 0.678 | 0.978 | 0.3 | <.01 | 1082 (774-1390) | 22 | EastSicilian | Welsh | – | – | – |
| Armenian | 16 | 0.76 | 0.946 | 0.83 | <.01 | 970 (494-1222) | 48 | Bulgarian | Turkish | – | – | – |
| Yi | 10 | 0.34 | 0.906 | 0.56 | <.01 | 1390 (858-1670) | 34 | Naxi | Han | – | – | – |
| Colombian | 7 | 0.902 | 0.881 | 0.36 | <.01 | 1782 (1670-1866) | 19 | Maya | Maya | – | – | – |

**Table S12:** Inferred dates (+ 95% CI; B=BCE, otherwise CE), proportions of admixture from minority source 1 (%) and the single present-day population that best matches each admixing source group, for all inferred events in the "full analysis". For the first event in "one date, multiway" and "multiple date" populations, for each source we provide both the best match and (in brackets) the population most represented in the inferred copying vector relative to the other source. For the second event in these cases, we only provide the population most represented in the inferred copying vector relative to the other source. $n_k$ is the number of sampled individuals; $R_1 = \max_i R_1^i$ as described in Note S4.8; "2E" and "MW" give $p$-values for multiple-dates and multiway admixture, respectively, as described in Note S4.6; and $FQ_B \equiv \min(FQ_1 + FQ_2, FQ_1^{\mathrm{NULL}} + FQ_2^{\mathrm{NULL}})$ is defined in Notes S4.5,S4.6 and used for calling events as "uncertain". †Dates/proportions/sources for these populations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these populations were inferred using the "multiple-date" grid; see Note S4.3.1. ΔThese populations also have "regional analysis" results; see Table S16.

### S6.3.2 additional events: Africa

Within sub-Saharan Africa, particularly sparse sampling makes some results difficult to interpret, though admixture is identified in all sub-Saharan African groups (Table S12) (but is uncertain for both Pygmy groups, and the Hadza who have a very small sample size of three) and is typically inferred to be complex (in five of eight characterised cases). Taking account of the fact that including closely related populations may mask subtle signatures of admixture, we designed several regional analyses (Note S7), including an Ethiopian-focused analysis described in Note S7.2 and a San-specific analysis described in Note S7.5.

Both Ethiopian populations show evidence of admixture occurring at multiple dates. In the full analysis, the most recent admixture in both the Ethiopian Jewish and Ethiopian populations date to within the broad period 1026-1670AD, covering part of the Zagwe and Solomonic dynasties, involving groups related to the modern day Egyptians, Saudi and Middle Eastern groups, and in both cases the "other" Ethiopian group. When we mask the "other" Ethiopian group from the analysis, the Ethiopian group involved in each event is replaced by the Bantu Kenyans (see Note S7.2), and estimated admixture dates do not strongly change. The earlier event in the full analysis of both populations spans a large period from 3930-10BCE and involves a similar Middle Eastern-like source (Saudi) one side and an Ethiopian-like source on the other. Given difficulties in dating admixture events at multiple times involving similar sources, either continuous admixture or multiple pulses of admixture over a very broad time range, are a possibility to explain our "multiple dates" signals here. We discuss these events further in Note S7.2.

The Bantu speaking populations share a strong genetic relationship with each other (69), as shown by the almost ubiquitous contribution of the Yoruba – genetically closely related to the sampled Bantu speaking groups, and from Nigeria – to all other African populations. This result is consistent with the expansion of Bantu-speaking peoples from the area of Nigeria/Cameroon (70), inferred to have resulted in DNA contribution using previous approaches (24; 63), and places an upper bound on their arrival in different regions. Genetic contributions into the Kenyan Bantus from sources inferred as related to the nearest sampled populations (Ethiopians, Sandawe) date to 1220CE (1050-1360CE), while a genetic contribution from a group related to the South African San (Khomani) to South African Bantus dates to the same period (1080-1360CE). The latter Khomani contribution to the South African Bantus is supported by an ADMIXTURE (7) plot in the analysis of (63) (see their Supplementary Figure S8), and our ADMIXTURE analysis, at a similar proportion to that we infer here (27%), though the authors did not attempt to date or describe this event.

We infer a single event in the West African Mandenka dating to 1390CE (1170-1590CE), between a sub-Saharan African group that largely includes the Yoruba and a second group also largely containing the Yoruba, but with additional more northern and eastern African composition – thus our inference of the source groups here does not allow clear interpretation. This date coincides with the development of extended exchange networks connecting West Africa and the Arab world (71). The mixture that we infer may therefore relate to the putatively increased trade and movement undertaken in the area during this time.

### S6.3.3 additional events: Europe

In general, we do not detect admixture in any of the Northwest European populations, and we discuss the results of our two regional analyses of the Mediterranean in Note S7.3 and Eastern Europe in the main paper and above.

Of the remaining European populations, the North Italians, Tuscans, Finnish and Norwegians show evidence of admixture, although all four are among the lowest scoring signals ($\max_i R_1^i < 0.7$; see Note S4.8).

In the Norwegians, we see evidence for two events occurring at a similar time, involving three groups. The three groups seem to include a majority group with a Northwest European haplotype set (related to Scottish or Irish), a more Russian-like group (identified by sharing haplotypes with East Asians, e.g. Yakut), and an Eastern European-like group (sharing haplotypes with e.g. Lithuanians), with the clearest dating signal coming from the East Asian-like haplotypes present. The events date to a period, 746CE (354-1166CE), that spans the Viking Age (but also overlaps a period with multiple episodes of migration in Europe). The Finns have only two samples, so results and dates are extremely tentative, but interestingly show a signal that is in some ways similar to that of Norway, but involving a far greater component of East-Asian-like, and Russian-like haplotypes, perhaps reflecting their geographic location as well as different history.

The North Italians exhibit a particularly early event, inferred as simple, which we date to 542BCE (1886BCE-606CE) using the "ancient" grid (see Note S4.3.1; date is estimated as 66BCE (776BCE-550CE) using the standard "recent" grid). The event involves a large amount of admixture (27-33%) between a population inferred as containing haplotypes similar to those of modern day West Asian (Cypriot) or Middle Easterners (Jordanians), and a population of Northwest European origin, inferred as most similar to present-day Welsh. We also note that the inferred event in the Tuscans involves a highly similar group of populations on both sides of the event (with Cypriot and French best representing each admixing source), but where the "Cypriot" fraction is greater (45%) and which we date to a more recent period 942CE (522-1222CE), although the confidence intervals for both dates do (barely) overlap.

### S6.3.4 additional events: Asia

The Northern Han Chinese (but not the Southern Chinese Han) show evidence of a small admixture event (6%), between an East-Asian Tujia-like group and, interestingly, a group similar to present-day populations from Armenia and Turkey, and dominated by haplotypes shared with modern-day West Eurasians, and discussed above.

The Xibo also show evidence of complex admixture occurring at multiple dates. The most recent involves Central Asian (Uzbekistani) groups on the one side (with some European-like haplotypes), and East Asian (Mongola-like haplotypes) on the other, dating very recently: to 1810CE (1698-1894CE). This date aligns well with the displacement of the Xibo from north-eastern China in 1764CE, during the expanding Qing dynasty in the 16-18th centuries (31), to the current location of the (HGDP) Xibo, who were sampled in Xinjiang on the western edge of China, towards Central Asia. We also see evidence for an older event occurring between local Oroqen and Uygur-like groups and Chinese-like (Naxi) groups occurring at 886CE (156-1166CE), so predating this displacement, and although the groups involved are consistent with the location of ancestors of the Xibo, the wide date range makes precise interpretation difficult.

The Yakut, hunter-gatherers from northern Siberia, show evidence of two different events occurring at different times. The first, that we date to 1754CE (1642-1866) with a proportion of $\approx$11%, appears to involve a mainly Russian-like group on one side, admixing with an Oroqen-like Siberian group on the other. This almost certainly relates to the first arrival of ethnic Russians in the Yakut heartlands, and annexation of Yakutia in the 1620s (67). The earlier event has a much broader range (634CE (710BCE-998CE)) and appears to have occurred between similar groups, but with a greater inferred Central Asian (Uzbek) element to the Russian-like source. We do not have a specific explanation for this event, but as the Yakut were renowned as fur-traders, this event could be related to movements associated with trade. However, as with Ethiopia, the involvement of similar (though perhaps not identical) groups for both events might also be consistent with this signal reflecting more continuous admixture – spanning our estimated dates – or equally with multiple admixture "pulses".

The results from our small sample of three Burmese individuals from Myanmar indicate a complex event occuring in 1306CE (1026-1726CE) between a Southeast Asian source with DNA similar to present-day Han, Cambodians and Dai, and a second source fit as a mixture of very widely dispersed present-day groups – the Han from China, Pathan from Pakistan, and Papuan from Papua New Guinea – suggesting this second source is not highly similar to any of the sampled populations. The Southeast Asian group may represent the Shans, a Tai people (closely related to the Dai) that migrated into present-day Burma as part of a larger Tai migration around the 11th or 12th century, and rapidly accelerated their migration during the Mongol invasions that overran the Pagan Empire in the late 13th century (72), potentially explaining our date estimate. The second possible event is more difficult to understand, involving a similar Melanesian-like group and another most closely related to the Colombians.

A simple event is seen in the Indians, though with a comparatively wide date range (438CE (150BCE-830CE)), and involves admixture between one population (86%) carrying haplotypes seen in a fairly small geographic region of northern Indian / Pakistan-like populations and another population (14%) with haplotypes widely dispersed across Central and Eastern Asia, and suggesting a group different from the other, all but one non-Indian, sampled groups. Our analysis of the Indian Jewish sample is suggestive of a complex event within a non-overlapping more recent time range involving West Asian and Middle Eastern-like (e.g. Iranian) sources on the one hand and Pakistan-like (e.g. Sindhi) sources on the other, which we date to 1362CE (1166-1530CE). The presence of a much more Middle Eastern and even European-like source in this event, particularly as it differs from the non-Jewish Indian event, might support the claim that some of the ancestors of this population migrated from the Middle East during the medieval period (23).

Iran shows a complex admixture signal, with evidence of a single date at 1306CE (1138-1474CE) ocurring at a very low proportion (2%). As noted in the main text, the most prominent event clearly involves (East) African ancestry and may be associated with the Arab Slave Trade. The second event shows admixture from a very different source, with components similar to DNA from a variety of present-day populations to the east of Iran, including haplotypes carried by e.g. Indian, Pakistan and Northeast Asian people. The date and – to an extent – the inferred sources of this admixture are consistent with (but need not imply with certainty) a DNA contribution from migrations related directly, or indirectly, to the Mongol Empire, as is seen in other regional groups sampled from within the historical Mongol Empire. (We do not include Iran in our list of seven groups with more strong evidence of a Mongol influence.)

Finally, we see an event in the Melanesians dating at a large proportion (49%). In our original analysis, this is dated to 1138CE (718-1502CE), but in our "NULL" analysis we note that although not inconsistent with this range, the date estimate differs substantively at 522CE (66BCE-1026CE), and from visual inspection of the coancestry curves (see Note S6.4 for more details) we suspect the older date may be more accurate. This event appears to have involved a local, islander Papuan-like group on the one hand (with much smaller contributions from diverse Asian sources, suggesting other haplotypes than those in Papuans were also carried by this source), and on the other a Southeast Asian group source inferred as best represented by Myanmar, and with strong contributions from e.g. the Han and other sources.

## S6.4 Robustness check of results

We tested the robustness of our model to each of the following potential issues:

1. varying levels of painting accuracy across genome, and other regional effects
2. variability in fine-scale recombination rates across human populations
3. choice of phasing protocol

4. value used for CHROMOPAINTER's inferred genome-wide average switching rate (i.e. $N_e$; see Appendix A.3)

For each of 1-4, we provide results for nine populations in Table S13. These populations represent a diverse range of admixture scenarios, including simple events, multiple sources (Maya, Kalash) and multiple dates (San Khomani), with varying admixture proportions (3-37%) and times (e.g. 598BCE in the Kalash to 1670CE in the Maya). These nine populations also represent a variety of inferred admixing source groups, illustrating admixture within Africa (South African Bantu, San Khomani), within Europe (Greek), within Asia (Cambodian, Turkish) and between sources from different continents (Druze, Kalash, Maya, San Khomani, West Sicilian).

In the final part of this section below, we also discuss consistency, in particular of admixture date inference, using our standard and "NULL" coancestry curves across all our sampled populations. This is motivated particularly by our finding in Note S5.5.3 that very strong bottlenecks following relatively old admixture may affect inference, and also by the fact that if our approach is working satisfactorily, the two types of curve ought to produce dates that approximately agree.

### S6.4.1 consistency for even/odd chromosome choice

It is possible that some parts of the genome are painted more accurately than others by CHROMOPAINTER, for example because some regions are genotyped more accurately than others or have been subjected to stronger selective effects (such as the Major-Histocompatibility Complex on chromosome 6; (73)). As our model's inference depends on averaging information across entire genomes, we believe it is unlikely to be strongly influenced by any particular genetic region(s), especially as correlations attributable to linkage disequilibrium (LD) typically do not extend beyond a few hundred kilobases in humans (74) and few regions are thought to be subject to strong selection.

Nonetheless to test for any such effect in practice, we analysed the even and odd chromosomes separately for each of the nine populations in Table S13. For these analyses, we generated copying vectors for each donor and recipient group using only the even or odd chromosomes, and generated curves for the nine recipient groups using only the even or odd chromosomes. We then used these copying vectors and curves to infer dates, proportions and admixing sources in these nine populations as described in Note S4. Results for the odd and even chromosomes (Table S13) are very similar and closely match that of our "full analysis" of Table S12, suggesting any systematic variation in painting accuracy does not have a strong effect, as predicted. Notably, differences in admixture results tend to be greater for those populations with more uncertain admixture dates as determined by the relevant 95% CIs, e.g. the Greek population, while results are very consistent for groups with narrow date ranges, e.g. the Druze, as expected. This property is repeated across the other robustness checks.

### S6.4.2 robustness to choice of genetic map

Recombination rates are known to vary among different human populations, though most of this variation is attributable to so-called recombination "hotspots" that are typically 1-2 kilobases in length (75; 76; 77; 78; 79; 46). At broader mega-base scales, i.e. scales that reflect the signals of admixture we attempt to characterize in this paper, recombination rates have been shown to be in strong agreement across continental human populations (46). Further the genetic map we used for our analyses was based on inferred rates averaged across the three different continental groups of Hap Map Phase 2, which would mitigate any differences among world-wide populations.

Nonetheless to test our findings' robustness to variable recombination rates, we re-analysed each of the nine populations in Table S13 using two additional, recently inferred genome-wide recombination maps. One map was generated by studying pedigrees in Icelandic individuals (54), while the other was generated by inferring crossover events in African-American individuals (46). For these analyses, we generated curves for the recipient groups using each of these different genetic maps, and then used these curves to infer dates, proportions and admixing sources as described in Note S4. When using the map of (54), we removed regions of our data that extended beyond the physical range of the genetic map, resulting in a slightly reduced set of 455,215 SNPs. Results for each genetic map are very similar and closely match that of our "full analysis" results in Table S12. This suggests that variable recombination rates among populations do not have a strong effect, as predicted.

### S6.4.3 robustness to phasing protocol

As CHROMOPAINTER uses haplotype information to infer relatedness among individuals, phasing is an important component of analysis. In theory, our inference may be biased if different groups within a dataset are phased differently, e.g. if only a subset are phased together. Specifically, phasing populations in such a manner would likely make them look more similar to one another than to other groups that were phased separately, and CHROMOPAINTER might therefore capture this false strong "relatedness". We believe in general that provided samples are phased "exchangeably", many details of phasing approach (e.g. the inclusion of panels of individuals with fixed phase assumed known, not subsequently included in the analysis) ought not to have a very strong impact. This is because equally shared noise or biases introduced by phasing choices should be accounted for by the mixture representation and painting "cleaning" step (see Notes S3.1-S3.3).

To avoid potential bias, we therefore phased all of our sampled individuals used in this study simultaneously (see Note S6.1). To improve phasing accuracy, we included Hap Map Phase 3 samples as reference haplotypes, as recommended by the authors of IMPUTEv2 (25). While there is no reason to suspect this might bias inference, again as all samples were phased together, we nonetheless re-phased our data using an alternative program SHAPEIT (52; 53) without including any reference haplotypes. In particular we phased each whole chromosome individually and used the default SHAPEIT values for all MCMC settings and input parameters, incorporating the same build 36 genetic map from Phase 2 of the HapMap project as in our initial phasing. We then generated copying vectors for each of the 95 populations using CHROMOPAINTER, generated curves for each of the nine recipient groups, and used these to infer dates, proportions and admixing sources as previously described. As predicted, results (Table S13) were very similar to our "full analysis" results in Table S12.

### S6.4.4 robustness to variation in CHROMOPAINTER's average "switch-rate"

We also tested the effect of CHROMOPAINTER's inferred "switch-rate" parameter (i.e. $N_e$ of Appendix A.3), which controls the average size of genetic segments that a recipient haplotype copies from a set of donors under the CHROMOPAINTER model (see Appendix A). In the analyses we present here, we infer this switch-rate using our data, via an Expectation-Maximization (EM) approach as described in e.g. Note S4.1.

To test GLOBETROTTER's robustness to switch-rate inference, we re-painted the chromosomes from each of the nine populations of Table S13 using our EM estimated value of switch-rate multiplied by a constant. We tried two different constants: 10 and 1/10. I.e. we re-painted chromosomes using a switch-rate ten times higher and ten times smaller than that used in our "full analysis" results of Table S12. We then generated curves for the nine recipient

groups using the painting results from each, and used these curves to infer dates, proportions and admixing sources as described in Note S4. Results are provided in Table S13.

When the switch-rate is ten times smaller, results are very similar to our previous findings. Results are also similar when the switch-rate is ten times higher, though not quite as concordant. In particular while inferred dates are very similar, inferred source groups sometimes vary. This likely reflects a decrease in power – increasing the switch-rate (in the extreme case, to infinity) drives our model towards that of approaches that consider each SNP independently during analysis and thus ignore haplotype information (e.g. ADMIXTURE (7), ROLLOFF (27)). Our results in (8) and in Notes S5.3 and S6.5 of this supplement suggest that this can result in a significant loss of power over our usual haplotype-based approach. Nonetheless, results are largely in agreement, regardless of whether the switch-rate used is ten times higher or ten times lower than that inferred by CHROMOPAINTER.

#### S6.4.5 consistency of coancestry curves with "NULL" coancestry curves

As described in Note S5.5.3, a strong bottleneck following an admixture event may result in our method slightly underestimating the date of admixture, with the effect becoming evident for admixture >20 generations ago. Standardizing our coancestry curves by those of a "NULL" individual, constructed to eliminate spurious signals of admixture (see Note S4.7), fixes this problem and provides reliable date estimates. However, inference of source groups may suffer under this latter approach. More generally, we expect both approaches to generate coancestry curves that largely agree in their conclusions, in particular their date estimates, and disagreement might signal lower reliability of the results. For these reasons, we decided to classify as "uncertain" any admixture scenarios where the 95% confidence intervals of the date estimates under the usual and "NULL"-standardizing approaches do not overlap. In the real data analysis, among 80 populations for which we identify admixture, there are three populations that are classified as "uncertain" due to this criterion: Biaka Pygmy, Mbuti Pygmy, and Japanese (see Table S12), while confidence intervals overlap in the remaining 77 cases.

There are two other cases where dates under the two scenarios differ enough that it may substantially affect interpretation of the events, even though the 95% CIs overlap. These are the Melanesian and the Namibian San, the latter only when analysed using the "San" regional analysis described in Note S7.5. For Melanesian, our standard analysis infers a date of 28 generations (95% CI: 15-43), while the "NULL" analysis infers a date of 50 generations (95% CI: 32-71). For the Namibian San, our standard analysis infers a date of 19 generations (95% CI: 15-23), while the "NULL" analysis infers a date of 44 generations (95% CI: 21-64).

For the Namibian San, we suspect the date under the "NULL" analysis is more accurate in this setting; this is the date closest to dates inferred by our standard (48 generations) and "NULL" analyses (40 generations) when performing the "full analysis" described in this Note (e.g. see Table S12). Visual inspection of the coancestry curves suggest our fitted date (green lines) under the "San" analysis is not a great fit to the data (black lines). For the Melanesian group, visual inspection of the coancestry curves similarly suggests that our fit is underestimating the true date of admixture reflected in the exponential decay curves. This again highlights the importance of visually inspecting the coancestry curves for all populations to see if GLOBETROTTER's default settings are adequately capturing the admixture signal in the data. The coancestry curves for all real data analyses in this paper are provided on-line at http://admixturemap.paintmychromosomes.com/.

We note that the analyses of all five populations discussed here have relatively weak signals of admixture according to our model. For example, all five analyses have a coefficient-of-determination ($R_1$) score below 0.86 (see Tables S12 and S16), which is lower than that of two-thirds of all 80 real data populations that we conclude admixture in for the "full analysis"

(see Table S12). This suggests that, in addition to the potential problem outlined here, our inference for these five groups might be affected by one or a combination of the following that would increase our uncertainty: low sample size, complex admixture, admixture between genetically similar groups that are difficult to distinguish, and a lack of relevant sampled groups to represent one or more of the admixing sources.

| | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Population** | $n_k$ | **analysis** | $R_1$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| BantuSA[1] | 8 | main | 0.955 | 0.12 | 0.47 | 1222 (1082-1362) | 27 | SanKhomani | Yoruba | – | – | – |
| | | odd | 0.931 | 0.46 | 0.55 | 1250 | 29 | SanKhomani | Yoruba | – | – | – |
| | | even | 0.935 | 0.89 | 0.52 | 1138 | 28 | SanKhomani | Yoruba | – | – | – |
| | | decode | 0.957 | 0.28 | 0.61 | 1194 | 27 | SanKhomani | Yoruba | – | – | – |
| | | af-amer | 0.964 | 0.96 | 0.55 | 1250 | 28 | SanKhomani | Yoruba | – | – | – |
| | | re-phase | 0.938 | 0.19 | 0.49 | 1166 | 28 | SanKhomani | Yoruba | – | – | – |
| | | switchx10 | 0.958 | 0.13 | 0.46 | 1166 | 26 | SanKhomani | Yoruba | – | – | – |
| | | switch/10 | 0.956 | 0.68 | 0.55 | 1194 | 27 | SanKhomani | Yoruba | – | – | – |
| Cambodian[1] | 10 | main | 0.919 | 0.22 | 0.07 | 1362 (1194-1502) | 19 | Uygur | Han | – | – | – |
| | | odd | 0.848 | 0.48 | 0.08 | 1418 | 13 | Pathan | Han | – | – | – |
| | | even | 0.859 | 0.51 | 0.08 | 1362 | 15 | Uygur | Han | – | – | – |
| | | decode | 0.914 | 0.31 | 0.1 | 1390 | 21 | Uygur | Han | – | – | – |
| | | af-amer | 0.906 | 0.31 | 0.1 | 1390 | 17 | Uygur | Han | – | – | – |
| | | re-phase | 0.906 | 0.28 | 0.11 | 1334 | 21 | Uygur | Han | – | – | – |
| | | switchx10 | 0.803 | 0.07 | 0.28 | 1446 | 16 | Myanmar | Han | – | – | – |
| | | switch/10 | 0.883 | 0.75 | 0.1 | 1362 | 10 | Pathan | Han | – | – | – |
| Druze[1] | 42 | main | 0.988 | 0.43 | 0.09 | 886 (774-998) | 3 | Yoruba | Cypriot | – | – | – |
| | | odd | 0.98 | 0.57 | 0.14 | 914 | 4 | Yoruba | Iranian | – | – | – |
| | | even | 0.984 | 0.78 | 0.18 | 914 | 3 | Yoruba | Iranian | – | – | – |
| | | decode | 0.988 | 0.35 | 0.12 | 858 | 2 | Yoruba | Cypriot | – | – | – |
| | | af-amer | 0.991 | 0.26 | 0.08 | 970 | 2 | Yoruba | Cypriot | – | – | – |
| | | re-phase | 0.987 | 0.36 | 0.12 | 886 | 3 | Yoruba | Iranian | – | – | – |
| | | switchx10 | 0.99 | 0.28 | 0.11 | 914 | 4 | Ethiopian | Iranian | – | – | – |
| | | switch/10 | 0.988 | 0.64 | 0.2 | 970 | 2 | Yoruba | Cypriot | – | – | – |
| Greek[1] | 20 | main | 0.823 | 0.8 | 0.34 | 914 (718-1138) | 37 | Polish | Cypriot | – | – | – |
| | | odd | 0.467 | 0.79 | 0.3 | 970 | 32 | Cypriot | Romanian | – | – | – |
| | | even | 0.662 | 0.87 | 0.14 | 522 | 33 | Cypriot | Romanian | – | – | – |
| | | decode | 0.719 | 0.75 | 0.37 | 830 | 46 | Cypriot | Hungarian | – | – | – |
| | | af-amer | 0.773 | 0.82 | 0.22 | 578 | 46 | Cypriot | Hungarian | – | – | – |
| | | re-phase | 0.742 | 0.57 | 0.15 | 466 | 32 | Syrian | Romanian | – | – | – |
| | | switchx10 | 0.798 | 0.15 | 0.35 | 662 | 33 | EastSicilian | Bulgarian | – | – | – |
| | | switch/10 | 0.683 | 0.82 | 0.24 | 578 | 46 | Cypriot | Hungarian | – | – | – |
| Kalash[M] | 23 | main | 0.778 | 0.16 | 0.01 | 598B (990-206B) | 35 | GermanyAustria [Scottish] | Pathan [Indian] | – | Naxi | Burusho |
| | | odd | 0.637 | 0.51 | <0.01 | 514B | 45 | Sindhi [Indian] | Turkish [Welsh] | – | Yi | Burusho |
| | | even | 0.597 | 0.66 | 0.04 | 626B | 44 | Tuscan [Scottish] | Pathan [Indian] | – | Naxi | Sindhi |
| | | decode | 0.702 | 0.37 | <0.01 | 626B | 45 | Tuscan [Scottish] | Pathan [Indian] | – | Naxi | Burusho |
| | | af-amer | 0.754 | 0.84 | <0.01 | 458B | 42 | Tuscan [Scottish] | Pathan [Indian] | – | Yi | Sindhi |
| | | re-phase | 0.725 | 0.75 | 0.01 | 934B | 40 | Bulgarian [GermanyAustria] | Pathan [Indian] | – | Yi | Sindhi |
| | | switchx10 | 0.779 | 0.68 | <0.01 | 430B | 38 | Iranian [Scottish] | Pathan [Burusho] | – | Burusho | Indian |
| | | switch/10 | 0.619 | 0.56 | <0.01 | 402B | 50 | Sindhi [Indian] | WestSicilian [Scottish] | – | Naxi | Burusho |
| Maya[M] | 21 | main | 0.994 | 0.28 | 0.01 | 1670 (1642-1726) | 19 | Spanish [Basque] | Pima [Colombian] | – | Yoruba | Karitiana |
| | | odd | 0.988 | 0.02 | 0.12 | 1698 | 19 | Spanish [Basque] | Pima [Colombian] | – | Hadza | Pima |
| | | even | 0.986 | 0.04 | 0.06 | 1670 | 18 | Spanish [Basque] | Pima [Colombian] | – | Hadza | Karitiana |
| | | decode | 0.994 | 0.6 | 0.13 | 1670 | 19 | Spanish [Basque] | Pima [Colombian] | – | Hadza | Pima |
| | | af-amer | 0.995 | 0.35 | 0.1 | 1670 | 18 | Spanish [Basque] | Pima [Colombian] | – | Hadza | Karitiana |
| | | re-phase | 0.99 | 0.63 | 0.1 | 1670 | 26 | Spanish [Basque] | Pima [Colombian] | – | Yoruba | Karitiana |
| | | switchx10 | 0.996 | 0.54 | 0.08 | 1670 | 14 | Egyptian [Hadza] | Pima [Surui] | – | Yoruba | Oroqen |

Continued on next page

| Table S13 – continued from previous page | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | first event | | | | second event signal | | |
| **Population** | $n_k$ | **analysis** | $R_1$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| | | switch/10 | 0.991 | 0.33 | 0.05 | 1698 | 18 | Spanish [Basque] | Pima [Colombian] | – | Yoruba | Karitiana |
| SanKhomani[2] | 30 | main | 0.991 | 0.01 | <0.01 | 1754 (1698-1838‡) | 27 | Egyptian [Welsh] | BantuSA [SanNamibia] | 1494B (10762B-970‡) | SanNamibia | BantuSA |
| | | odd | 0.984 | 0.01 | <0.01 | 1782 | 36 | BantuSA [Welsh] | SanNamibia [SanNamibia] | 662 | SanNamibia | BantuSA |
| | | even | 0.986 | 0.02 | <0.01 | 1726 | 37 | BantuSA [Welsh] | SanNamibia [SanNamibia] | 1186B | SanNamibia | BantuKenya |
| | | decode | 0.991 | 0.01 | <0.01 | 1754 | 33 | BantuSA [Welsh] | SanNamibia [SanNamibia] | 262B | SanNamibia | BantuKenya |
| | | af-amer | 0.989 | <0.01 | <0.01 | 1754 | 30 | Egyptian [Welsh] | BantuSA [SanNamibia] | 66B | SanNamibia | BantuSA |
| | | re-phase | 0.987 | 0.01 | <0.01 | 1726 | 32 | BantuSA [Scottish] | BantuSA [SanNamibia] | 1158B | SanNamibia | BantuSA |
| | | switchx10 | 0.971 | 0.01 | <0.01 | 1754 | 35 | BantuSA [IndianJew] | SanNamibia [SanNamibia] | 430B | SanNamibia | BantuKenya |
| | | switch/10 | 0.985 | 0.01 | <0.01 | 1754 | 28 | French [Welsh] | SanNamibia [SanNamibia] | 1102B | SanNamibia | BantuSA |
| Turkish[1] | 17 | main | 0.985 | 0.08 | 0.26 | 1250 (1166-1362) | 8 | Mongola | Iranian | – | – | – |
| | | odd | 0.966 | 0.04 | 0.32 | 1222 | 8 | Mongola | Iranian | – | – | – |
| | | even | 0.982 | 0.54 | 0.25 | 1250 | 8 | Mongola | Iranian | – | – | – |
| | | decode | 0.985 | 0.24 | 0.32 | 1222 | 7 | Mongola | Iranian | – | – | – |
| | | af-amer | 0.987 | 0.18 | 0.24 | 1278 | 7 | Mongola | Iranian | – | – | – |
| | | re-phase | 0.98 | 0.1 | 0.31 | 1138 | 8 | Mongola | Iranian | – | – | – |
| | | switchx10 | 0.976 | 0.05 | 0.11 | 1250 | 17 | Uzbekistani | Iranian | – | – | – |
| | | switch/10 | 0.981 | 0.37 | 0.27 | 1250 | 8 | Mongola | Iranian | – | – | – |
| WestSicilian[1] | 10 | main | 0.941 | 0.07 | 0.13 | 1166 (914-1362) | 4 | Yoruba | EastSicilian | – | – | – |
| | | odd | 0.895 | 0.11 | 0.12 | 1390 | 3 | Yoruba | EastSicilian | – | – | – |
| | | even | 0.888 | 0.63 | 0.08 | 858 | 3 | Yoruba | EastSicilian | – | – | – |
| | | decode | 0.941 | 0.06 | 0.11 | 1306 | 5 | Yoruba | EastSicilian | – | – | – |
| | | af-amer | 0.928 | 0.03 | 0.08 | 1278 | 5 | Yoruba | EastSicilian | – | – | – |
| | | re-phase | 0.92 | 0.1 | 0.07 | 1278 | 2 | Yoruba | EastSicilian | – | – | – |
| | | switchx10 | 0.954 | 0.42 | 0.14 | 1222 | 5 | Egyptian | EastSicilian | – | – | – |
| | | switch/10 | 0.922 | 0.07 | 0.17 | 1306 | 3 | Yoruba | EastSicilian | – | – | – |

**Table S13:** Results demonstrating robustness of our "full analysis" inference for nine populations. The "analysis" column denotes the analysis used, with "main" our reported inference from Table S12, "odd" the inference when analysing odd chromosomes only, "even" analysing even chromosomes only, "decode" using the genetic map from deCode (54), "af-amer" using the African-American genetic map of (46), "re-phase" based on a re-phasing of the data using SHAPEIT (52; 53), "switchx10" based on increasing CHROMOPAINTER's switch-rate parameter (i.e. $N_e$ of Appendix A.3) by a factor of 10, and "switch/10" based on decreasing CHROMOPAINTER's switch-rate parameter by a factor of 10. Other columns are labeled as in Table S12. Our model's inferred conclusion for each population in the main "full analysis" is given next to its name in the first column, with $^1$= "one-date", $^M$= "one date, multiway", and $^2$= "multiple-dates". ‡Dates in "main" results for these populations were inferred using the "multiple-date" grid; see Note S4.3.1.

## S6.5 Comparison to other approaches

### S6.5.1 ADMIXTURE

By way of comparison to previous methods, we used ADMIXTURE (7) to visualise admixture components in our full dataset of 1530 individuals, using different values of user-supplied clusters $K$. As recommended by the authors, we thinned the dataset to remove SNPs in linkage disequilibrium by recursively removing SNPs within a 500kb sliding window, pruning SNPs where the $r^2$ correlation in allele frequencies across the dataset was greater than 0.4. This resulted in a pruned dataset containing 143,204 SNPs.

We ran ADMIXTURE in unsupervised mode for all values of $K \in [2, ..., 14]$; applying the recommended cross-validation procedure gave $K = 10$ as the optimal solution. To visualise the results, within each labeled population we computed the average of the ancestry components across all of that populations' indidviduals. We present these population-averaged results for $K \in [8, ..., 14]$ in Figure S15.

We note at this point that because ADMIXTURE and GLOBETROTTER have different aims and properties, we cannot completely verify our results in this manner. ADMIXTURE infers the mutation frequencies of sampled groups as a mixture of some fixed number $K$ of components, inferred from the data. These components provide strong geographic information in many cases, at the continental level (because the number of components that can be meaningfully analysed is usually <20). The inference of a group as a mixture of these components can be the result of recent admixture, but can also be produced by more ancient admixture (e.g. older than that we can infer), or continuous migration and/or spatial structure, which we again did not generally infer as an admixture signal in our simulations. Thus, we do not expect all groups inferred as a mixture in the ADMIXTURE analysis to necessarily show a GLOBETROTTER signal. Conversely, the inference of a group or set of groups as being almost entirely from a single component indicates genetic drift in these group(s), which might reflect lengthy isolation indicating no recent admixture, but might also reflect a strong relatively recent bottleneck subsequent to such an admixture event, and so in some cases we might observe a GLOBETROTTER signal in such cases. Nevertheless, we believe that in *most* cases where only a single component is seen we do not expect to see a GLOBETROTTER signal, and otherwise there ought to be a likelihood of strong recent genetic drift. Conversely, where there is a strong recent admixture event inferred by GLOBETROTTER - between groups different enough to be separable (for large enough $K$) by ADMIXTURE - we would expect to see some signal in the ADMIXTURE results.

Clearly, ADMIXTURE also does not attempt to finely spatially resolve genetic contributions, or estimate times of ADMIXTURE events, and so we are unable to test these aspects of our results. For the latter, we compare our results to those of ROLLOFF in the next section.

When tested against these criteria, in general there is close agreement between the ADMIXTURE components and our GLOBETROTTER analysis. The conclusion from our results that most worldwide populations appear recently admixed to some degree is clearly consistent overall with the ADMIXTURE plots.

Looking at $K = 10$, around 12 groups are roughly a single colour "block" in the ADMIXTURE inference, and GLOBETROTTER does not identify and characterize admixture in 9 of the 12 groups. One exception are the Kalash, who we believe have experienced strong drift since admixture, perhaps leading to their population-specific group. Indeed for $K = 9$, the Kalash have a signal of European-related ancestry consistent with the type of signal we observe, though perhaps at a lower proportion of only around 10%. The other exceptions are the Mandenka, where GLOBETROTTER's inferred admixture is between two groups both appearing closely similar to the Yoruba, and thus likely not separable by ADMIXTURE, and the SanNamibia, whose inferred admixture fraction is just 2% in the GLOBETROTTER analysis,

and who also show the weakest admixture signal (measured by coefficient of determination, 0.318) among all the events that we identified and characterised. Conversely, among those groups with no evidence of admixture ($p \geq 0.01$), excluding groups we infer to be admixed but record as "uncertain", only a group of five European populations (five from the British Isles, and GermanyAustria) and the Tujia are represented as a clear mixture in the ADMIXTURE analysis. The European groups show almost indistinguishable ADMIXTURE results to one another, suggesting any event(s) may have influenced them all, and thus is most likely to be ancient. This might also be the case for the Tujia, or their admixture history might be masked by the inclusion of the genetically similar Han, a potential problem we highlighted in the main text.

Of the populations where GLOBETROTTER does infer admixture between different groups (e.g. groups in different continental regions), the genome-wide summaries provided by the ADMIXTURE analysis are typically consistent with the GLOBETROTTER analysis. This is the case for the events discussed in the main text – e.g. a small Northeast Asian contribution is seen specifically in East European groups but is tiny or absent in many other European groups. Sub-Saharan African admixture into the seventeen populations bordering the Mediterranean Sea, Arabian Sea and the Persian Gulf (with admixture inferred by both approaches to potentially correspond most closely to North African groups in the case of the S. Italians, E. Sicilians, and Moroccans) is clearly seen in this set of groups. European, North African, Middle Eastern or West Asian related admixture is inferred by GLOBETROTTER in ten diverse groups excluding these regions themselves and Central Asia: the Indian Jewish, Tu, Han N. China, Mongola (as a second event signal), Yakut, Maya, Pima, San Khomani and both Ethiopian groups. This is seen, with similar source regions, in all ten cases in the ADMIXTURE analysis. The only other two groups with a similar ADMIXTURE signal are the Sandawe (GLOBETROTTER results suggest a Middle Eastern component may actually have come indirectly, through a group related to the Ethiopians) and Hadza (GLOBETROTTER is uncertain about the nature of admixture in this group, with only 3 samples).

Considering more subtle events, these are more difficult to compare because ADMIXTURE does not always distinguish the inferred source groups. However, notable is that a dark blue East Asian component appears in the ADMIXTURE analysis only at $K \geq 10$. ADMIXTURE finds this component to be most strongly represented in the present-day Yakut. Within Asia, this component is prominent in the northern Chinese / Siberian populations, in all of which GLOBETROTTER also finds evidence of introgression from sources related to the Yakut or Oroqen. This element is also present in Central Asian populations, tapering off to the west. Using GLOBETROTTER, we find the strongest signal of admixture in these Central Asian populations to be a "Mongol-like" signal (i.e. consistent with the time of the Mongol Empire), a subtlety that appears consistent with results from ADMIXTURE (given the roughly 50% dark blue and purple mix in each case) but which ADMIXTURE cannot strongly determine at even $K = 14$ clusters.

### S6.5.2 ROLLOFF

We also ran ROLLOFF (27) on all 30 populations in which GLOBETROTTER concluded a single admixture event between only two source groups in the "full analysis" (i.e. all the populations under "one date" in Table S12). As in our simulations (see Note S5.3), we ran ROLLOFF to generate and fit the LD decay curve from 1 to 50cM using a grid bin-size of 0.1cM, otherwise using default parameters. To calculate standard errors for ROLLOFF's date estimates, we used the Weighted Block Jackknife procedure described in (50), removing one chromosome at a time, as recommended by the authors (27).

For each of these 30 populations, we ran ROLLOFF using the sampled populations listed in

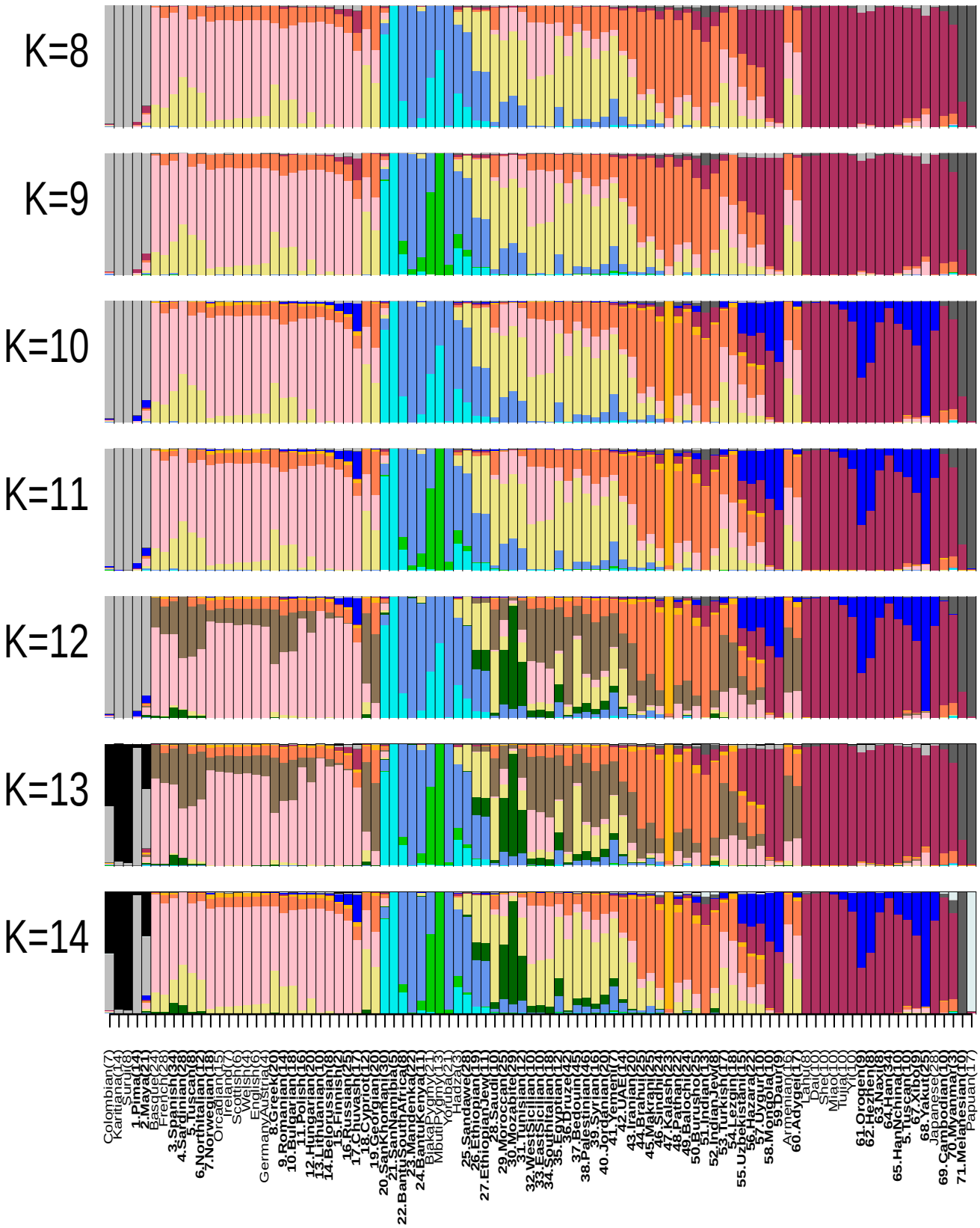



**Figure S15:** Colors represent each of $K$ distinct clusters inferred by ADMIXTURE (7) applied to our full dataset, with colors chosen to match (roughly) those used in Figures S12 and S13 for the fineSTRUCTURE clades from Table S11. Each column gives the proportion of genome-wide DNA assigned to each cluster, averaged across individuals in each of the 95 populations labeled at bottom (number of individuals from each in parentheses). Population labels in bold and numbered are ones where GLOBETROTTER inferred adxmixture, with the numbers referring to the position on the world map in Figure 2D of the main text.

Table S14 as surrogates to represent each admixing source group. For most populations, these surrogates match the sampled groups that GLOBETROTTER inferred to best represent the admixing sources for that population, as given in Table S12. However, ROLLOFF failed to run without error for populations in which at least one of the surrogates had fewer than 10 sampled individuals. Therefore, in such cases we re-ran ROLLOFF using a similar population that did have at least 10 sampled individuals. (It is possible that adding additional, genetically similar individuals to each surrogate would improve estimation for ROLLOFF, though we note that (27) used 10 as a minimum number of individuals per surrogate when inferring admixture in similar data using related, drift-based techniques.) For Mandenka, GLOBETROTTER infers Yoruba to best represent both admixing source groups; for this population we therefore ran ROLLOFF using Yoruba and Moroccan as surrogates, selecting the latter based on GLOBETROTTER's inferred mixing coefficients.

The inferred dates (in generations) for GLOBETROTTER and ROLLOFF are provided in Table S14 for these 30 populations. For one population (Georgian), ROLLOFF falied to run without error, though we tried two different sets of donors (Adygei and Turkish, Adygei and Greek). For most of the remaining 29 populations, the dating estimates from ROLLOFF validate those inferred by GLOBETROTTER. For example, the one standard error region around GLOBETROTTER's estimated date overlaps with the one standard error region around ROLLOFF's estimated date in 24 of the 29 cases (83%), with the regions around two standard errors overlapping in all 29 cases. However, we note that ROLLOFF's date estimates had standard errors above 40 for 10 populations. GLOBETROTTER also found it relatively challenging to infer an admixture signal in these 10 populations (based on their relative goodness-of-fit $R_1$ values – see Table S12), presumably due to issues such as small sample sizes in the admixed population and/or the admixing source groups being genetically similar. In such scenarios, our simulations predict GLOBETROTTER to have increased power to identify admixture over approaches like ROLLOFF (see Note S5.3). We note that our standard errors are relatively small for even these groups, and that GLOBETROTTER has a lower standard error in all 29 populations.

| Population | $n_k$ | source 1 | source 2 | ROLLOFF | GLOBETROTTER |
|---|---|---|---|---|---|
| Hazara | 22 | Mongola (10) | Iranian (13) | 23 ± 1 | 22 ± 0.9 |
| Uzbekistani | 15 | Mongola (10) | Iranian (13) | 20 ± 1.4 | 19 ± 1.1 |
| Uygur | 10 | Mongola (10) | Iranian (13) | 23 ± 2.6 | 22 ± 1.3 |
| Makrani | 22 | BantuKenya (11) | Balochi (21) | 18 ± 1.8 | 18 ± 1.2 |
| Druze | 42 | Yoruba (21) | Cypriot (12) | 39 ± 7.3 | 37 ± 1.9 |
| Mozabite | 25 | Yoruba (21) | Moroccan (22) | 23 ± 1.9 | 21 ± 1.3 |
| Turkish | 17 | Mongola (10) | Iranian (13) | 28 ± 3.2 | 24 ± 1.5 |
| Brahui | 23 | BantuKenya (11) | Balochi (21) | 13 ± 3.4 | 20 ± 1.5 |
| Yemeni | 4 | BantuKenya (11) | Syrian (16) | 15 ± 2.3 | 14 ± 1.8 |
| Pima | 14 | Turkish (17) | Maya (21) | 9 ± 3.6 | 6 ± 0.9 |
| BantuSA | 8 | SanKhomani (30) | Yoruba (21) | 26 ± 2.5 | 25 ± 2.3 |
| Tu | 10 | Greek (20) | HanNchina (10) | 33 ± 6.3 | 25 ± 2.3 |
| WestSicilian | 10 | Yoruba (21) | EastSicilian (10) | 26 ± 7.8 | 27 ± 3.9 |
| Cambodian | 10 | Uygur (10) | Han (34) | 17 ± 4.7 | 20 ± 2.7 |
| Georgian | 20 | Adygei (17) | Greek (20) | – | 30 ± 3.3 |
| Romanian | 13 | Lithuanian (10) | EastSicilian (10) | 85 ± 52.7 | 31 ± 2.6 |
| Bulgarian | 18 | Polish (16) | Cypriot (12) | 102 ± 51.6 | 28 ± 3.5 |
| Hezhen | 8 | Tujia (10) | Mongola (10) | 12 ± 256.8 | 13 ± 1.3 |
| Oroqen | 9 | Yakut (25) | Mongola (10) | 16 ± 266.4 | 15 ± 2 |
| Hungarian | 18 | Cypriot (12) | Polish (16) | 65 ± 24 | 39 ± 3.5 |
| HanNchina | 10 | Turkish (17) | Tujia (10) | 37 ± 11.1 | 26 ± 3.8 |
| Daur | 9 | Tujia (10) | Mongola (10) | 59 ± 40.7 | 21 ± 1.7 |
| Greek | 20 | Polish (16) | Cypriot (12) | 69 ± 18.5 | 36 ± 3.7 |
| Melanesian | 10 | Papuan (16) | Cambodian (10) | 66 ± 12.1 | 28 ± 7.6 |
| Mandenka | 22 | Moroccan (22) | Yoruba (21) | 22 ± 10.3 | 19 ± 4.2 |
| Indian | 13 | Cambodian (10) | Sindhi (23) | 91 ± 41.1 | 53 ± 8.4 |
| NorthItalian | 12 | Cypriot (12) | French (28) | 157 ± 317.6 | 71 ± 11.8 |
| Polish | 16 | French (28) | Lithuanian (10) | 826 ± 569.1 | 31 ± 5.1 |
| Tuscan | 8 | Cypriot (12) | French (28) | 165 ± 198.7 | 35 ± 6.1 |
| SanNamibia | 5 | Sandawe (28) | SanKhomani (30) | 62 ± 187.4 | 48 ± 8.9 |

**Table S14:** Comparison between date estimates of GLOBETROTTER (in generations, with standard deviations across 100 bootstrap re-samples) and ROLLOFF (with standard errors), for all populations for which GLOBETROTTER concludes a single admixture event between two source groups in the "full analysis" (i.e. "one date" in Table S12). $n_k$ is the number of sampled individuals for the putatively admixed population. "source1" and "source2" list the surrogates used in ROLLOFF to represent the two admixing groups, with the number of sampled individuals from each surrogate in parentheses. Populations are ordered as in Table S12. ROLLOFF failed to run without error for the "Georgian" population.

# S7 Additional "regional" analyses of sample collection

As noted in Note S4.1, we exclude self-copying when painting individuals' chromosomes in order to prevent the masking of admixture signals common to all individuals with the same population label, i.e. to prevent individuals copying only from their own population at all genomic locations. However, this does not account for the possibility that samples with two different population labels share a very similar admixture history, which will similarly attenuate the signal of such an event(s). For this reason, in addition to the "full analysis" described in Note S6.3 we performed the slightly modified analyses shown in Table S15 that removed specific donor populations from each analysis, informed in part by our fineSTRUCTURE analysis of Note S6.2. We refer to these as the "regional analyses".

Additional populations were also excluded as potential donors for these populations, most often for computational reasons. For example, Native American and Oceania populations are rarely copied from by populations outside of their own geographical regions, and so were

| Analysis Name | Recipient Pops | Donor Pop exclusions |
|---|---|---|
| "**Middle East + North Africa**" | Bedouin, Egyptian, Iranian, Jordanian, Moroccan, Mozabite, Palestinian, Syrian, Tunisian, UAE | Americas, Oceania |
| "**Ethiopian**" | Ethiopian, Ethiopian Jew | Americas, Oceania |
| "**Mediterranean**" | East Sicilian, Greek, Sardinian, South Italian, Spanish, West Sicilian | Americas, Middle East + North Africa, Oceania |
| "**Central Asia**" | Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi, Uygur, Uzbekistani | Americas, Oceania |
| "**San**" | San Khomani, San Namibia | Americas, Oceania |
| "**East Europe I**" | Belorussian, Bulgarian, Chuvash, Finnish, Hungarian, Lithuanian, Polish, Romanian, Russian | Africa, Americas, Southeast Asia, Oceania |
| "**East Europe II**" | Belorussian, Bulgarian, Chuvash, Finnish, Hungarian, Lithuanian, Romanian, Russian | Africa, Americas, Southeast Asia, Oceania |

**Table S15:** Details of "regional analyses". Each row represents a distinct analysis, where individuals from each population in the second column were allowed to copy from all donor populations in the dataset excluding the populations listed in the second and third columns. In the third column, "Africa" refers to all samples from the populations [Bantu Kenya, Bantu South Africa, Biaka Pygmy, Egyptian, Ethiopian, Ethiopian Jew, Hadza, Mandenka, Mbuti Pygmy, Sandawe, San Namibia, San Khomani, Yoruba], "Americas" refers to all samples from the populations [Colombian, Karitiana, Maya, Pima, Surui], "Southeast Asia" refers to all samples from the populations [Cambodian, Dai, Han, Han N.China, Japanese, Lahu, Miao, Myanmar, Naxi, She, Tu, Tujia, Xibo, Yi], and "Oceania" refers to all samples from the populations [Melanesian, Papuan].

excluded as donors for all recipient populations in Table S15. Similarly the "East Europe" populations did not show any signs of admixture from sub-Saharan Africa or from Southeast Asian populations in the "full analysis"; hence populations from these geographic regions were excluded as donors in the "East Europe I" and "East Europe II" analyses.

In each of these analyses, we did not allow either the recipient or remaining donor individuals to copy from any recipient individuals when generating their copying vectors. This differs from the "full analysis", where donors and recipients were allowed to copy from a subset of recipient individuals (and for example "self-copying" was allowed in the recipients). The copying vectors for donor populations were generated allowing each donor group to copy from all but one individual of each other donor group, mimicking the "full analysis" (see Note S4.1.1). However, to generate the recipient groups' copying vectors and painting samples, each recipient individual was allowed to copy from all individuals of each donor population. These steps were done for computational reasons and to e.g. ensure that the protocol for making copying vectors for all donor individuals within an analysis matched. As a very similar protocol was followed for our "real-sample" simulations (see Note S5.1.5), we do not expect this slight change to affect inference substantially. Another minor difference from the "full analysis" is that after using CHROMOPAINTER's E-M algorithm to jointly estimate $N_e$ and $\theta$ for each individual as described in Note S4.1.2, we then averaged the inferred values across all individuals from the region, rather than for each population separately, before performing our final CHROMOPAINTER run. For example, for the "East Europe I" analysis, we averaged estimated $N_e$ and $\theta$ across all individuals from the "East Europe I" populations, rather than using distinct values of $N_e$ and $\theta$ for each "East Europe I" recipient population in our final

CHROMOPAINTER run. This was done to ensure that any differences we observed among populations within a region was not attributable to the (usually very slight) variations in our model's inferred $N_e$ and $\theta$. Otherwise the protocol followed that outlined in Note S4.

We briefly describe aspects of each "regional analysis" below, and summarize these results in Table S16, Figures S16-S21 and Figures 3 and 4 of the main text. For each regional analysis, we provide figures (e.g. Figure S16) depicting the differences between the inferred copying vectors of the two admixing sources for every inferred event, on a blue-red scale using values defined in Note S4.8, together with our estimated date(s) of admixture.

We present results for all analyses, including the "full analysis" that uses all populations as potential donors (see Note S6), in the on-line interactive map at
http://admixturemap.paintmychromosomes.com/.

## S7.1 "Middle East + North Africa"

Because we observed similar admixture in many of our near Eastern groups, and these groups were difficult for fineSTRUCTURE to separate, we performed an "Middle East + North Africa analysis" to clear up admixture signals, for example from African sources, that may be obscured in the "full analysis". To represent a local source, we retained the Saudi as a potential donor, as this population had less evidence of African admixture than the other Middle East and North Africa populations in the "full analysis" results (African admixture into the Saudi Arabians is inferred only as the second strongest signaled event in the full analysis). Here and for the remainder of this Note we compare our "regional" and "full" analysis results for consistency of events detected.

Our "Middle East + North Africa" analysis shows, in all 10 groups analysed (Figure S16), admixture within the past millenium involving a sub-Saharan African source (always inferred West African, except in the UAE and Iranian samples where it is "BantuKenya"), and a West Eurasian source, always involving an excess of copying from the Saudi group, and inferred to share the largest number of haplotypes (among the allowed donors) with either the Saudi population, or one of the most nearby included donors (e.g. the Cypriot population). These events, and donor groups, closely match those of the original "full" analysis (see main text) and all date confidence intervals overlap. Admixture fractions vary from 3% (in the Iranians) to 21% (in the Moroccans), again agreeing well with the full analysis (see main text).

Apart from these strongest events, we infer a complex history for seven groups, three with multi-way admixture and four with admixture at more than one time. Encouragingly, for six of these groups the conclusion is identical to that in the original "full" analysis. In these cases (although there are small differences in particular groups chosen within small geographical regions), dates and groups involved are generally highly consistent with the full analysis (e.g. dating confidence intervals always overlap). The only exception are the Tunisians, who almost but did not quite reach the cutoff (showing $p = 0.05$) for admixture at more than one time in the full analysis, and in this regional analysis show $p = 0.03$. Conversely in the "full" analysis, the Egyptians just fall below the cutoff (having $p = 0.05$) and are inferred to have two events, while they are inferred to have a single event time in the regional analysis. It is worth noting that in nine of the ten groups the p-value for "multiple dates" falls below 0.25, suggesting admixture may have affected these groups at multiple times in almost all cases, though we may not always have power with our sample sizes to detect this. Furthermore, in almost all cases the older event involves similar groups to the more recent event, consistent with more continuous admixture over a long time range (with the range 690-1750AD containing all but one of the date estimates). These dates are consistent with the slave trade (35), over an approximate time range (650-1900CE) that very closely corresponds to our observed range of estimated event times. The exception is the Moroccans, with an old inferred event involving

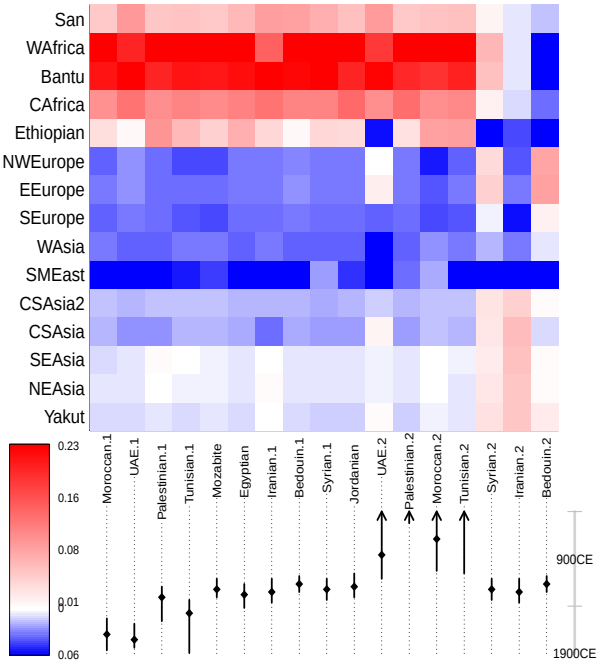



**Figure S16:** For each of 17 inferred events (x-axis) among 10 labeled populations from the Middle East and North Africa, analysed under the "Middle East + North Africa" protocol described in Note S7.1, this heatmap depicts the difference between the two admixing sources' inferred DNA proportions copied from each of 15 geographic regions on the y-axis. Geographic regions that are colored more red are inferred to strongly represent one admixing source group, while regions colored more blue strongly represent the other admixing source group. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively. As in all plots in this section, column labels suffixed with ".1" show the most recent or strongly signalled event for this group, while column labels suffixed with ".2" show the older or less strongly signalled event.

more European-like groups (e.g. the Basque from Spain) than the more recent event (which infers the Saudis as the differently "copied from" donor, along with European-like groups). While continuous migration similar to that inferred by (20) is plausible in the Moroccans, these results suggest additional events involving other groups may have also occurred.

For the groups with multi-way admixture, the Iranians show a signal very similar to the "full" analysis signal discussed above. The Syrians and the Bedouin both seem to have signals indicating 3-way admixture between sub-Saharan Africans, groups to their south (e.g. sharing haplotypes with the Saudis) and groups to their immediate north and west (e.g. sharing haplotypes with the Cypriots and Georgians), and with the nature of this event similar for the full and regional analyses. Finally, one difference with the "full" analysis is that we now fail to find evidence of multiple events in the Jordanians. Although we still find the African admixture event in this group, we no longer see evidence of a second event inferred to involve haplotypes shared with other Middle Eastern groups (excluded in the new painting) and groups to the north (e.g. sharing haplotypes with the Georgians) and dated to the range 1050-1330CE, perhaps because we do not allow these groups as donors in the regional analysis.

## S7.2 "Ethiopian"

We designed an Ethiopian-specific analysis that excluded the two Ethiopian populations from copying from each other. In this analysis, we find evidence for two events in both populations in agreement with the "full analysis". We infer one ancient (≈1000BCE, but with large confidence

intervals) and one much more recent event (≈1502CE and ≈1614CE) in these two populations. Interestingly, the effect of masking the "other" Ethiopian population allows us to pick very similar groups for both events in both populations. The most recent event involves a largely similar sub-Saharan African component admixing into an Egyptian-like group (Figure S19). The ancient event also involves a sub-Saharan African source admixing with a much more Eurasian source, that is most closely related to the modern day Saudis among our sampled groups. The complexity in composition of the sources for these admixture events perhaps reflects the various African and non-African groups (including Europeans and Arabs) that came into contact to control trade along the Red Sea and beyond, and are consistent with either overland or maritime migrations of people. These conclusions are almost identical to our "full" analysis findings, and dates also agree. Our conclusions of multiple dates imply that either two pulses of admixture from the direction of Eurasia, potentially including from the unsampled region to the south of Egypt, have affected this region, or that more continuous migration, over the past three to five millennia or more, has taken place (80; 81).

## S7.3 "Mediterranean"

Similar to the "Middle East + North Africa" analysis, some of the "Mediterranean" populations were difficult for fineSTRUCTURE to separate, and were frequently used as donors to one another in inferred admixture events. We also wished to investigate whether the admixture into these populations involves DNA contributions from a sub-Saharan African source versus from Middle Eastern (or North African) sources. For this reason, we performed a "regional" analysis and excluded Middle Eastern and North African populations as donors, apart from the Saudi population and the more genetically distinct Cypriots, as in the "Middle East + North Africa" analysis. Figure S17 shows the results of this analysis.

Results for West Sicily, Greece and Spain were almost unchanged in all respects from the "full" analysis, and suggest direct West African DNA dating to around 1220CE in the West Sicilians, the Greek signal discussed in the main text, and complex admixture at more than one time in the Spanish that is difficult to precisely characterise. Recent and ancient events in the Spanish both involve West African haplotypes, inferred (at least for the more recent event) to occur within a source population with additional ancestry related to present-day Europeans, and for the older event in the "full" analysis to also involve haplotypes carried by North Africans (the Mozabites), excluded in the regional analysis. The other inferred admixing group has haplotypes shared with present-day Spanish Basques. This signal suggests admixture might be continuous and involve interactions with north and sub-Saharan Africans over at least the past ≈2000 years. In addition, we see strong evidence ($p < 0.01$) that admixture also involves multiple distinct groups in both cases, implying yet more complexity in the genetic history of this Spanish sample, which our method is unable to interpret fully. For instance our curves (available for both types of analysis at http://admixturemap.paintmychromosomes.com/) show segments of DNA shared with Basque, Tuscan and Yoruban are all negatively correlated at short genetic distances, implying these three types of haplotypes are carried on different ancestral backgrounds, in addition to (and not captured by) our two-date inference.

For East Sicily, Sardinia and South Italy, a second event was inferred by either the full or regional analysis.

In East Sicilians, both analyses show a recent event whose age overlaps the period 1050-1390CE, and involving admixture from a source group carrying west African haplotypes (Yoruba, Mandenka) with a southern-European like group (sharing haplotypes with North Italian and Tuscan individuals). Both analyses show very similar evidence ($p = 0.05$) of two admixture dates, but this is just below our threshold only in the "full" analysis. In this case, the older event occurs between somewhat similar sources (though haplotypes carried by Bantu speakers

from Kenya are inferred as most over-represented, rather than West African haplotypes), and the signal dates back at least to 186AD, pre-dating the Arab slave trade. This might reflect continuous-type admixture, or might (given the separation in inferred dates) reflect a distinct earlier event involving a population with some African ancestry.

In Sardinians, admixture is inferred at exactly the same time (630AD) in both analyses, and there is one very similar inferred event involving one population from North Africa or the Mediterranean. Specifically, this group is inferred as most similar to Cyprus, but with additional West African haplotypes, in the regional analysis, and as showing high levels of haplotype sharing with each of sub-Saharan (east) Africa, North Africa, and the Middle East in the full analysis – note North Africa and the Middle East cannot act as donors in the "Mediterranean" analysis. The other group is inferred to be from Europe, sharing haplotypes with people from Wales, Italy, France and Spain. In the full analysis, a second event ($p = 0.01$), is inferred involving admixture between groups sharing haplotypes with populations to the west of Sardinia (e.g. Basque) and populations around the northern Mediterranean coast and to the east (including South Italian, and Cypriot). Evidence of a second event is seen in the regional analysis ($p = 0.07$) but does not reach our significance threshold, perhaps because of a lack of informative included donors.

Finally, a somewhat similar signal is seen in South Italians, for a recent event inferred by both methods, involving one group inferred as genetically most similar to Cyprus or Egypt, and characterised by carrying haplotypes shared with West African (Mandenka) people, within a wide overlapping date range. Both analyses show evidence of multi-way admixture ($p < 0.05$), but the "Mediterranean" analysis also shows a signal (marginally significant at $p = 0.05$) of admixture at multiple times involving similar groups. Therefore this potentially reflects long-term admixture, similar to the multiple-date signal seen in the East Sicilians. The "multi-way" admixture is characterised in the full analysis, and suggests another event involving a group sharing haplotypes with middle-eastern, Cyprus and Arabian peninsula populations (Yemeni, Cypriot, among others). Overall, both the the "full" and "Mediterranean" analyses suggest very complex genetic exchange across the Mediterranean over 2000 years or more, also involving people from sub-Saharan African groups, with multiple events inferred in 4 of 6 groups though with dates and details differing between groups.

## S7.4 "Central Asia"

Due to the relatively high density of populations sampled from Central Asia in the HGDP (22), we disallowed copying among populations sampled from within this region for this "Central Asia" analysis. We describe our interpretation of these results in detail in Note S6.3 and the main paper, including Figure 4, with analagous plots to those for other analyses in this section provided in Figure S18.

## S7.5 "San"

We describe the San Khomani results in detail in Note S6.3.1, and the results appear very similar in this analysis, the main difference being that the SanNamibia are replaced as a donor in the second event by a group inferred to share haplotypes with several other, predominantly African populations (e.g. individuals from Ethiopia, the Sandawe, but also some Middle Eastern groups).

By masking the San Khomani, we found evidence for complex admixture in the Namibian San (Figure S19). Our small sample size (5 individuals) makes details of the analysis uncertain, but we find an event dating to 1390CE (1280-1500CE). We note though that our "NULL" analysis and our original "full" analysis both give a somewhat earlier, though overlapping, date

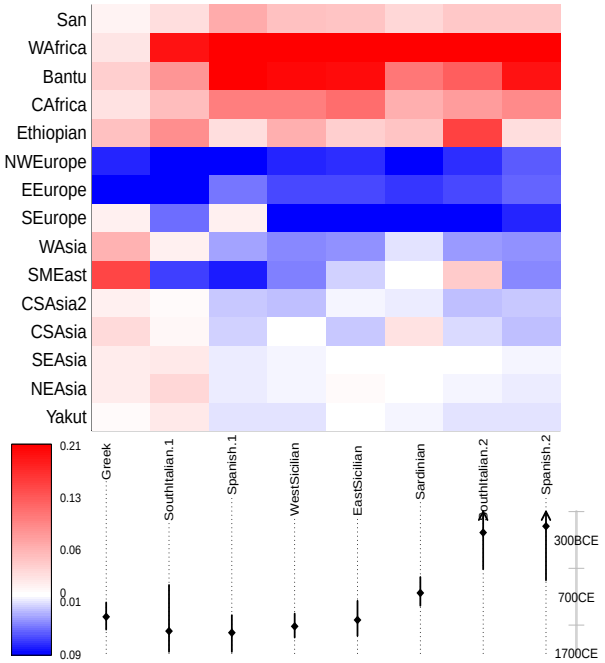



**Figure S17:** For each of 8 inferred events (x-axis) among 6 labeled populations along the Mediterranean Sea, analysed under the "Mediterranean" protocol described in Note S7.3, this heatmap depicts the difference between the two admixing sources' inferred DNA proportions copied from each of 15 geographic regions on the y-axis. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively.

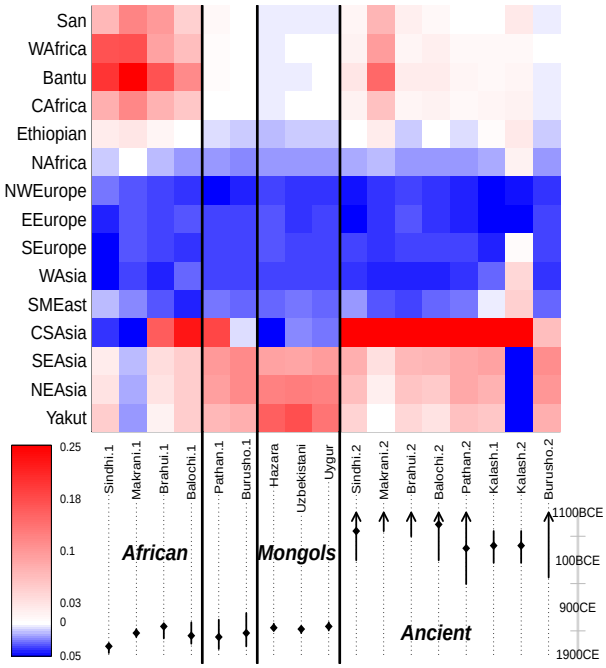



**Figure S18:** For each of 17 inferred events (x-axis) among 10 labeled populations from Central Asia, analysed under the "Central Asia" protocol described in Note S7.4, this heatmap depicts the difference between the two admixing sources' inferred DNA proportions copied from each of 15 geographic regions on the y-axis. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively.

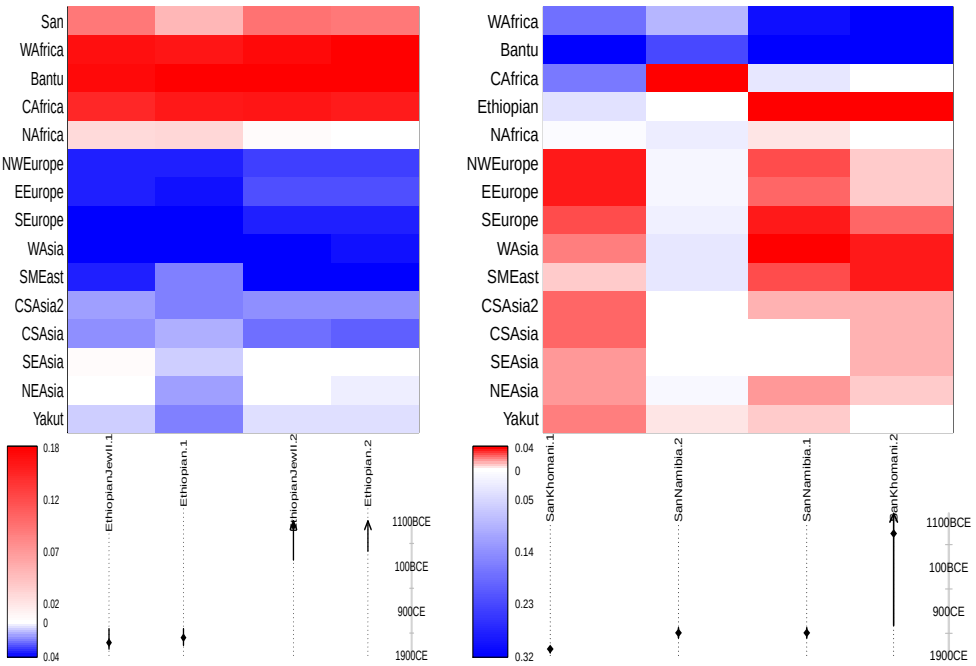



**Figure S19:** For each of 8 inferred events (x-axis) among 4 labeled populations from Ethiopia (left) and southern Africa (right), analysed under the "Ethiopian" and "San" protocols described in Notes S7.2 and S7.5, respectively, these heatmaps depict the difference between the two admixing sources' inferred DNA proportions copied from each of 15 geographic regions on the y-axis. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively.

range of 690CE (130CE-1334CE) which may be more reliable – see Note S6.4.5 for details.

One source group for the Namibian San is inferred to share haplotypes with multiple populations, including many groups outside Africa (e.g. the Cypriots and Saudi), Ethiopian groups within Africa, the Sandawe and northwards. Apart from the Sandawe, which has a weak signal, all of these groups share positive coancestry signals, implying copying from any of them can mark the same source. The other group shares haplotypes with all available African groups to the south of the Sandawe. This event, although strongly signalled, does not seem to have a straightforward historical interpretation. Our results might be consistent with admixture from a group related to Bantu speaking populations into an earlier population, at an uncertain location. A second but more weakly signaled event, perhaps less reliable given our small sample size, is inferred to involve a group widely sharing haplotypes with current-day populations in the same broad region, including the Hadza and the Mbuti Pygmies.

The Namibian San are in fact the only population in our regional analyses with an event time whose 95% interval does not (quite) overlap that of the "full" analysis (this original inferred event involved only 2% admixture and therefore might have been almost entirely masked by the San Khomani sharing). Moreover, and despite the strong sharing of haplotypes with our San samples, neither of the San Khomani events occur at an inferred time overlapping that of this event. Instead the most similar event in any other group within our data to that seen in the Namibian San occurs in the Sandawe, whose more recent event includes a similar set of groups on each side, dates to the same time, and involves a similar inferred ancestry fraction (37% vs. 41%), tentatively suggesting events of a similar nature occurring in these widely separated groups. We anticipate additional analyses of larger samples may help shed light on these puzzles in the future.

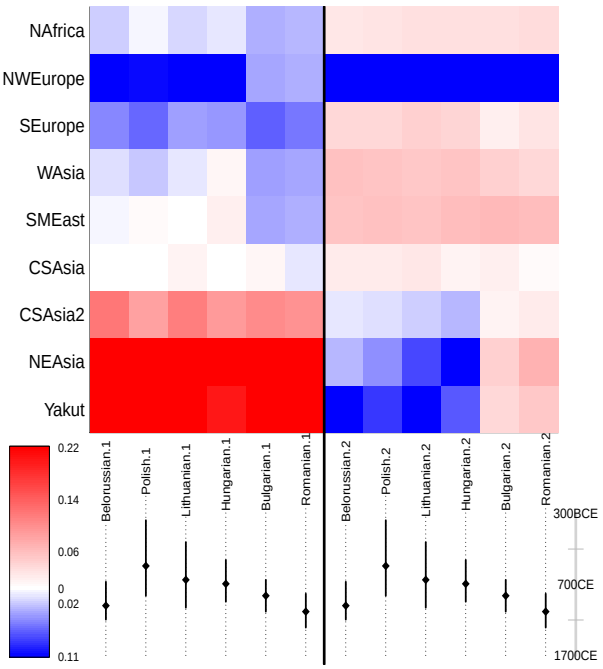



**Figure S20:** For each of 12 inferred events (x-axis) among 6 labeled populations from eastern Europe, analysed under the "East Europe I" protocol described in Note S7.6, this heatmap depicts the difference between the two admixing sources' inferred DNA proportions copied from each of 9 geographic regions on the y-axis. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively. Analogous to Figure S21, which gives results for analyses under the "East Europe II" protocol.

## S7.6 "East Europe"

The difference between the "East Europe I" and "East Europe II" analyses is that the latter analysis included the Polish as a potential donor population. The Polish were included in this analysis to reflect a Slavic language speaking source group. The results from "East Europe I" are summarized in Figure S20, with analogous results for "East Europe II" provided in Figure S21. We describe our interpretation of these results in detail in the main paper (see Figure 3 of the main text) and Note S6.3.

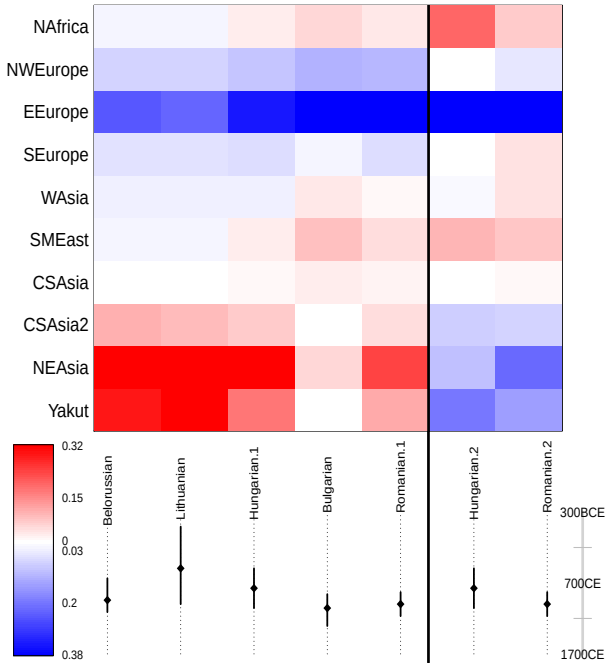



**Figure S21:** For each of 7 inferred events (x-axis) among 5 labeled populations from eastern Europe, analysed under the "East Europe II" protocol described in Note S7.6, this heatmap depicts the difference between the two admixing sources' inferred DNA proportions copied from each of 10 geographic regions on the y-axis. Solid diamonds and lines below each population give the inferred admixture date and 95% CI, respectively. Analogous to Figure S20, which gives results for analyses under the "East Europe I" protocol.

| | | | | | | first event | | | | second event signal | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Population** | $n_k$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| **MIDDLE EAST + NORTH AFRICA** | | | | | | | | | | | | |
| Jordanian[1] | 18 | 0.992 | 1 | 0.38 | 0.2 | 1194 (1054-1306) | 6 | Yoruba | Cypriot | – | – | – |
| Egyptian[1] | 10 | 0.989 | 1 | 0.21 | 0.36 | 1278 (1166-1418) | 14 | Yoruba | Cypriot | – | – | – |
| Mozabite[1] | 25 | 0.987 | 1 | 0.06 | 0.4 | 1222 (1110-1306) | 13 | Yoruba | EastSicilian | – | – | – |
| Syrian$^M$ | 16 | 0.987 | 0.999 | 0.21 | 0.01 | 1222 (1110-1334) | 5 | Yoruba [Yoruba] | Turkish [Cypriot] | – | Saudi | Yoruba |
| Bedouin$^M$ | 45 | 0.993 | 1 | 0.13 | 0.02 | 1166 (1082-1250) | 5 | Yoruba [Yoruba] | Saudi [Saudi] | – | Hadza | Saudi |
| Iranian$^M$ | 13 | 0.958 | 0.991 | 0.07 | 0.03 | 1250 (1110-1362) | 3 | BantuKenya [BantuKenya] | Turkish [Saudi] | – | Naxi | Saudi |
| Moroccan[2] | 22 | 0.974 | 1 | 0.01 | 0.27 | 1698 (1530-1866) | 21 | Yoruba [Yoruba] | EastSicilian [Saudi] | 690 (234B-1026) | Yoruba | Basque |
| Tunisian[2]‡ | 12 | 0.981 | 1 | 0.03 | 0.18 | 1474 (1334-1894) | 13 | Yoruba [Yoruba] | EastSicilian [Saudi] | 234B (2474B-1054) | Yoruba | Saudi |
| Palestinian[2]‡ | 46 | 0.994 | 1 | 0.03 | 0.19 | 1306 (1194-1558) | 8 | Yoruba [Yoruba] | Cypriot [Saudi] | 542B (3230B-522) | Yoruba | Cypriot |
| UAE[2]‡ | 9 | 0.98 | 1 | 0.04 | 0.07 | 1754 (1586-1838) | 4 | BantuSA [BantuKenya] | Saudi [Saudi] | 858 (5190B-1110) | BantuSA | Saudi |
| **ETHIOPIAN** | | | | | | | | | | | | |
| EthiopianJew[2]‡ | 11 | 0.894 | 1 | 0.02 | 0.25 | 1614 (1306-1754) | 43 | BantuKenya [Hadza] | Egyptian [Jordanian] | 990B (4546-234B) | Hadza | Saudi |
| Ethiopian[2]‡ | 19 | 0.937 | 1 | 0.04 | 0.23 | 1502 (1306-1670) | 27 | BantuKenya [Hadza] | Egyptian [Jordanian] | 1886B (3062-430B) | BantuKenya | Saudi |
| **MEDITERRANEAN** | | | | | | | | | | | | |
| EastSicilian[1] | 10 | 0.941 | 1 | 0.05 | 0.12 | 1110 (774-1390) | 4 | Yoruba | Tuscan | – | – | – |
| WestSicilian[1] | 10 | 0.941 | 1 | 0.12 | 0.12 | 1222 (998-1418) | 3 | Yoruba | Tuscan | – | – | – |
| Sardinian[1] | 28 | 0.919 | 0.999 | 0.42 | 0.07 | 634 (354-858) | 9 | Cypriot | French | – | – | – |
| Greek[1] | 20 | 0.777 | 0.999 | 0.1 | 0.12 | 1054 (802-1278) | 37 | Polish | Cypriot | – | – | – |
| Spanish[2]‡ | 34 | 0.97 | 0.999 | 0.02 | <.01 | 1334 (1026-1670) | 14 | French [Yoruba] | French [Basque] | 542B (5666B-410) | Yoruba | Basque |
| SouthItalian[2]‡ | 18 | 0.926 | 0.999 | 0.05 | 0.04 | 1306 (494-1670) | 26 | Cypriot [Mandenka] | Tuscan [Welsh] | 430B (6534B-214) | Mandenka | Welsh |
| **CENTRAL ASIA** | | | | | | | | | | | | |
| Hazara[1] | 22 | 0.996 | 1 | 0.59 | 0.37 | 1334 (1250-1390) | 47 | Mongola | Turkish | – | – | – |
| Uzbekistani[1] | 15 | 0.993 | 1 | 0.25 | 0.4 | 1362 (1306-1446) | 38 | Mongola | Turkish | – | – | – |
| Uygur[1] | 10 | 0.99 | 1 | 0.36 | 0.5 | 1306 (1222-1390) | 50 | Mongola | Turkish | – | – | – |
| Kalash$^M$† | 23 | 0.854 | 0.992 | 0.6 | 0.01 | 570B (1186B-18) | 22 | Indian [Indian] | Turkish [Lithuanian] | – | Chuvash | Indian |
| Kalash$^M$ | 23 | 0.828 | 0.994 | 0.32 | 0.02 | 402B (710-38B) | 27 | Indian [Indian] | Turkish [Lithuanian] | – | Naxi | Indian |
| Sindhi[2]‡ | 23 | 0.852 | 0.996 | <.01 | 0.01 | 1726 (1670-1866) | 11 | Egyptian [Yemeni] | Iranian [Lezgin] | 710B (3958-94B) | Indian | Lithuanian |
| Brahui[2]‡ | 23 | 0.967 | 0.999 | 0.02 | <.01 | 1306 (1306-1558) | 4 | BantuSA [SanNamibia] | Iranian [Lezgin] | 3678B (9894-598B) | Indian | Lezgin |
| Balochi[2]‡ | 21 | 0.94 | 0.997 | 0.02 | 0.03 | 1502 (1222-1670) | 8 | Indian [Indian] | Iranian [UAE] | 850B (8550-94B) | Indian | Lezgin |
| Makrani[2]‡ | 22 | 0.991 | 1 | 0.02 | 0.21 | 1446 (1362-1530) | 4 | BantuSA [Hadza] | Iranian [IndianJew] | 3034B (5890-710B) | Indian | Lezgin |
| Pathan[2]‡ | 22 | 0.938 | 1 | 0.03 | 0.11 | 1530 (1166-1782) | 37 | Turkish [Indian] | Iranian [Yemeni] | 346B (7262B-410) | Indian | Lithuanian |
| Burusho[2]‡ | 25 | 0.968 | 1 | 0.04 | 0.23 | 1446 (1026-1726) | 22 | Tu [Yi] | Turkish [IndianJew] | 2446B (4602B-270) | Yi | Lezgin |
| **SAN** | | | | | | | | | | | | |
| SanNamibia$^M$ | 5 | 0.776 | 0.995 | 0.57 | 0.01 | 1390 (1278-1502) | 41 | BantuSA [EthiopianJew] | BantuSA [BantuSA] | – | Hadza | BantuSA |
| SanKhomani[2]‡ | 30 | 0.992 | 1 | 0.01 | 0.32 | 1754 (1726-1838) | 18 | Turkish [Scottish] | BantuSA [BantuSA] | 850B (4238B-1222) | EthiopianJew | BantuSA |
| **EAST EUROPE I** | | | | | | | | | | | | |
| Chuvash[1] | 16 | 0.958 | 1 | 0.15 | 0.48 | 886 (690-1054) | 28 | Mongola | GermanyAustria | – | – | – |
| Finnish[1] | 2 | 0.559 | 1 | 0.64 | 0.33 | 298 (514B-1054) | 9 | Oroqen | Norwegian | – | – | – |
| Finnish[1]† | 2 | 0.551 | 1 | 0.87 | 0.33 | 690 (346B-1558) | 11 | Mongola | Norwegian | – | – | – |
| Polish$^M$ | 16 | 0.757 | 0.999 | 0.91 | <.01 | 438 (206B-858) | 4 | Yakut [Yemeni] | GermanyAustria [Norwegian] | – | Greek | Norwegian |
| Bulgarian$^M$ | 18 | 0.893 | 0.999 | 0.81 | <.01 | 858 (634-1082) | 4 | Mongola [Oroqen] | Greek [Greek] | – | Orcadian | Greek |
| Lithuanian$^M$ | 10 | 0.706 | 0.997 | 0.91 | <.01 | 634 (102-1026) | 10 | Uygur [Hezhen] | GermanyAustria [Norwegian] | – | Norwegian | Greek |
| Romanian$^M$ | 13 | 0.92 | 1 | 0.59 | <.01 | 1082 (830-1306) | 4 | Oroqen [Oroqen] | Greek [Greek] | – | Orcadian | Greek |

Continued on next page

| Table S16 – continued from previous page | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | first event | | | | second event signal | | |
| **Population** | $n_k$ | $R_1$ | $FQ_B$ | **2E** | **MW** | **date** | **%** | **source 1** | **source 2** | **date** | **source 1** | **source 2** |
| Hungarian$^M$ | 18 | 0.848 | 1 | 0.75 | 0.01 | 690 (354-942) | 16 | Turkish [Daur] | GermanyAustria [Welsh] | – | Greek | Norwegian |
| Belorussian$^M$ | 8 | 0.739 | 1 | 0.9 | 0.02 | 998 (662-1194) | 8 | Uygur [Yakut] | GermanyAustria [Norwegian] | – | Greek | Norwegian |
| Russian$^2$‡ | 25 | 0.961 | 1 | 0.05 | 0.37 | 1222 (1054-1614) | 10 | Mongola [Yakut] | GermanyAustria [Scottish] | 2082B (5302-542B) | Yakut | Norwegian |
| **EAST EUROPE II** | | | | | | | | | | | | |
| Russian$^1$ | 25 | 0.964 | 1 | 0.09 | 0.68 | 886 (774-1054) | 10 | Mongola | Polish | – | – | – |
| Bulgarian$^1$ | 18 | 0.88 | 1 | 0.52 | 0.05 | 1054 (858-1306) | 41 | Cypriot | Polish | – | – | – |
| Belorussian$^1$ | 8 | 0.759 | 1 | 0.99 | 0.55 | 942 (634-1110) | 2 | Oroqen | Polish | – | – | – |
| Lithuanian$^1$ | 10 | 0.73 | 1 | 0.95 | 1 | 494 (94B-998) | 1 | Oroqen | Polish | – | – | – |
| Finnish$^1$ | 2 | 0.579 | 1 | 0.95 | 0.09 | 298 (178B-942) | 9 | Oroqen | GermanyAustria | – | – | – |
| Romanian$^M$ | 13 | 0.926 | 1 | 0.68 | <.01 | 998 (830-1166) | 23 | Turkish [Hezhen] | Polish [Polish] | – | Hezhen | UAE |
| Hungarian$^M$ | 18 | 0.86 | 0.986 | 0.93 | <.01 | 774 (494-1054) | 7 | Uygur [Yemeni] | Polish [Polish] | – | Yemeni | Polish |
| Finnish$^M$† | 2 | 0.568 | 1 | 0.99 | 0.04 | 410 (850B-1390) | 10 | Oroqen [Oroqen] | Norwegian [Norwegian] | – | Orcadian | Adygei |
| Chuvash$^2$‡ | 16 | 0.956 | 1 | 0.04 | 0.58 | 1334 (1026-1754) | 33 | Mongola [Oroqen] | Polish [Polish] | 1046B (5218B-326) | Oroqen | Polish |

**Table S16:** Inferred dates (+ 95% CI; B=BCE, otherwise CE), proportions of admixture from minority source 1 (%) and the single present-day population that best matches each admixing source group, for all inferred events in the "regional analyses". For the first event in "one date, multiway" and "multiple date" populations, for each source we provide both the best match and (in brackets) the population most represented in the inferred copying vector relative to the other source. For the second event in these cases, we only provide the population most represented in the inferred copying vector relative to the other source. $n_k$ is the number of sampled individuals; $R_1 = \max_i R_1^i$ as described in Note S4.8; "2E" and "MW" give $p$-values for multiple-dates and multiway admixture, respectively, as described in Note S4.6; and $FQ_B \equiv \min(FQ_1 + FQ_2, FQ_1^{\text{NULL}} + FQ_2^{\text{NULL}})$ is defined in Notes S4.5,S4.6 and used for calling events as "uncertain". Our model's inferred conclusion for each population is given next to its name in the first column, with $^1$= "one-date", $^M$= "one date, multiway", and $^2$= "multiple-dates". †Dates/proportions/sources for these populations were inferred using the "ancient" grid; see Note S4.3.1. ‡Dates for these populations were inferred using the "multiple-date" grid; see Note S4.3.1.

# Appendices

## A Details of the painting algorithm

We used an approach based on a previously described Hidden Markov model (42) to perform an initial "chromosome painting" that forms the basis of our method to understand admixture events. In the setting of studying admixture, the aim of this approach is to identify, as we move along the genome, an approximation to the nearest relative of a genome potentially influenced by admixture, with these relatives identified among members of other sampled populations (8). Recombination means ancestral relationships, and so nearest relatives, change as we move along the genome, in a dynamic process that provides information on the underlying admixture structure. We take an approach we refer to as "chromosome painting", which pieces together each individual's genome as a block-by-block mosaic of the genomes of individuals sampled from other labeled populations. We describe this process in detail in sections A.1-A.6 below. The algorithm we use is all contained in a program called CHROMOPAINTER, which is first described in (8) and is freely available at www.paintmychromosomes.com.

### A.1 The Model

The Li and Stephens algorithm (42) is a likelihood based model that captures the principal features of the genealogical process with recombination while remaining computationally tractable for large datasets. Under the model, a chromosome is generated segment-by-segment by copying from a conditional set of fixed haplotypes. In our notation, every individual consists of two haploids, each consisting of a single phased haplotype per chromosome. The $L$ total SNPs in each haploid are listed one chromosome at a time, in order within each chromosome.

Suppose that we wish to generate a particular haploid $h_* = \{h_{*1}, ..., h_{*L}\}$, with $h_{*l}$ the observed allele of $h_*$ at site $l$, using $j$ pre-existing donor haploids $h_1, ..., h_j$. Let $\vec{\rho} = \{\rho_1, ..., \rho_{L-1}\}$ be a vector of genetic distances, with $\rho_l$ the population-scaled genetic distance between sites $l$ and $l+1$ (i.e. $\rho_l = N_e g_l$, where $N_e$ is a "recombination scaling constant" analogous to "effective population size", and $g_l$ is the genetic distance in Morgans between sites $l$ and $l+1$). (Between chromosomes, the genetic distance between the last site of the previous chromosome and the first site of the next chromosome is $\infty$.) Let $\vec{f} = \{f_1, ..., f_j\}$ be a vector of copying probabilities, with $f_k$ the probability of copying from haploid $h_k$ at any site. Let $\theta$ correspond to a per site mutation (or "imperfect copying") parameter. The conditional probability $\Pr(h_* \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta)$ is structured as a Hidden Markov Model (HMM). Let $\vec{X} = \{X_1, ..., X_L\}$ represent the hidden state sequence vector, with $X_l$ the existing haploid from the set $h_1$,...,$h_j$ that haploid $h_*$ copies from at site $l$. Switches in the haploid being copied between $X_l$ and $X_{l+1}$ occur as a Poisson process with rate $\rho_l$. The transition probabilities for $X$ between sites $l$ and $l+1$ are as follows (we exclude $h_1, ..., h_j$ and the parameters from the left-hand side conditional probabilities of Equations S26 and S27 below for ease of reading):

$$\Pr(X_{l+1} = x_{l+1} | X_l = x_l) = \begin{cases} \exp(-\rho_l) + \Big(1 - \exp(-\rho_l)\Big) f_{x_{l+1}} & \text{if } x_{l+1} = x_l; \\ \Big(1 - \exp(-\rho_l)\Big) f_{x_{l+1}} & \text{otherwise,} \end{cases} \tag{S26}$$

The observed state sequence component of the Hidden Markov Chain, the probability of observing a particular allele given the haploid that $h_*$ is copying from at a given SNP, allows for "imperfect" copying:

$$\Pr(h_{*l} = a | X_l = x) = \begin{cases} 1.0 - \theta & h_{xl} = a; \\ \theta & h_{xl} \neq a. \end{cases} \quad \text{(S27)}$$

Here $h_{kl}$ refers to the allelic type of haploid $k$ at SNP $l$. To calculate $\Pr(D) \equiv \Pr(h_* \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta)$, a summation is performed over all permutations of the copying process, i.e. a summation over all possible $x$, which can be accomplished efficiently using the forward algorithm (e.g. (82)).

For analyses presented here, we jointly estimate the recombination scaling constant $N_e$ and the mutation parameter $\theta$ using 10 steps of an Expectation-Maximization (E-M) algorithm as described below. Unless otherwise noted (e.g. Note S6.4.2), we fix each $g_l$ by taking the build 36 genetic distance estimates from the HapMap website (*http://www.hapmap.org/*), which were calculated using Phase II genotypes and averaging values across the three HapMap populations as described in (44). We also fix each $f_k$ to be $1/j$ for $k = 1, ..., j$, allowing for equal *a priori* probability of copying from each conditional haploid.

## A.2 Forwards and Backwards probabilities

Again excluding parameters from conditional probabilities for ease of reading, let

$$\alpha_{kl} = \Pr(h_{*1}, ..., h_{*l}, X_l = h_k)$$

denote the *forward probabilites* and

$$\beta_{kl} = \Pr(h_{*(1+1)}, ..., h_{*L} \mid X_l = h_k)$$

the *backward probabilities* of the HMM.

We calculate $\alpha_{kl}$ for $k = 1, ..., j$ in the following manner (82):

1. $\alpha_{k1} = \Pr(h_{*1} \mid X_1 = h_k) f_k$
2. $\alpha_{kl} = \Pr(h_{*l} \mid X_l = h_k) \Big( \big[ \sum_{i=1}^{j} \alpha_{i(l-1)} \big] f_k \big(1 - \exp(\rho_l)\big) + \exp(\rho_l) \alpha_{k(l-1)} \Big)$ for $l = 2, ..., L$.

We calculate $\beta_{kl}$ for $k = 1, ..., j$ in the following manner (82):

1. $\beta_{kL} = 1.0$
2. $\beta_{kl} = \big[ \sum_{i=1}^{j} \beta_{i(l+1)} f_i \Pr(h_{*(l+1)} \mid X_{l+1} = h_i) \big] \big(1 - \exp(\rho_l)\big) + \exp(\rho_l) \Pr(h_{*(l+1)} \mid X_{l+1} = h_k) \beta_{k(l+1)}$ for $l = 1, ..., (L-1)$.

## A.3 Using the E-M algorithm to estimate the scaling parameter $N_e$

One can take a fixed $N_e$ for calculating $\vec{\rho}$, or use the E-M to find a local maximum of $N_e$ in the following manner. Start with an initial value of $N_e$, and at each iteration of the E-M replace $N_e$ with:

$$N_e^* = \frac{\sum_{l=1}^{L-1} \big( [\sum_{k=1}^{j} \hat{f}_{k,l}] \rho_l / [1.0 - \exp(-\rho_l)] \big)}{\sum_{l=1}^{L-1} g_l}. \quad \text{(S28)}$$

Here $\hat{f}_{k,l} = \frac{1}{\Pr(D)} \Big[ \alpha_{k(l+1)} \beta_{k(l+1)} - \alpha_{kl} \beta_{k(l+1)} \Pr(h_{*(l+1)} | X_{l+1} = h_k) \exp(-\rho_l) \Big]$ and represents the probability haploid $h_*$ copies from haploid $h_k$ at site $l$ given at least one "switch" has occurred between $l-1$ and $l$. Each $\hat{f}_{k,l}$ and $\rho_l$ are calculated using the previous value of $N_e$. (For starting values in this paper, we take $N_e = 400,000/J$, with $J$ equal to the total number of donor and

recipient haploids analysed at a time in the "full analysis" of Note S6, or the total number of donor haploids analysed at a time in the "separate analyses" of Note S7 and simulations of Note S5.5.3. However, we note that results should converge to a similar value even if one used starting values within a factor of 10 or so of these.)

Unless otherwise mentioned, for analyses presented here we estimate $N_e$ (jointly with $\theta$) for each recipient population studied using 10 iterations of E-M for chromosomes 1, 4, 10, and 15. We then average estimated $N_e$ across these four chromosomes, weighting by chromosome length in cM, to get a final estimate of $N_e$ that we use for generating "painted chromosome" samples and "copying vectors".

## A.4 Using the E-M algorithm to estimate the mutation parameter $\theta$

One can take a fixed $\theta$ for calculating (S27), or use the E-M to find a local maximum of $\theta$ in the following manner. Start with an initial value of $\theta$, and at each iteration of the E-M replace $\theta$ with:

$$\theta^* = \frac{\sum_{l=1}^{L} \left( \sum_{i=1}^{j} \alpha_{il}\beta_{il} I_{[h_{*l} \neq h_{il}]} / \Pr(D) \right)}{L}. \tag{S29}$$

Here $I_{[h_{*l} \neq h_{il}]}$ is an indicator that the allele $h_{*l}$ carried by the recipient is not equal to allele $h_{il}$ carried by donor haploid $i$ at SNP $l$, and each $\alpha_{il}$, $\beta_{il}$ and $\Pr(D)$ are calculated using the previous value of $\theta$. (For starting values in this paper, we take $\theta$ equal to the value used in (42), i.e. $\theta = (1/2)(\sum_{i=1}^{J} 1/i)^{-1}/[j + (\sum_{i=1}^{J} 1/i)^{-1}]$ for $J$ total haploids – including all recipient and donor individuals – considered in the particular analysis.)

For analyses presented here, we estimate $\theta$ (jointly with $N_e$) for each recipient population studied using 10 iterations of E-M for chromosomes 1, 4, 10, and 15. We then average estimated $\theta$ across these four chromosomes, weighting by chromosome length in cM, to get a final estimate of $\theta$ that we use for generating "painted chromosome" samples and "copying vectors".

## A.5 "Copying vectors": calculating expected lengths of genome copied

To calculate $\hat{f}_1, ..., \hat{f}_j$, the posterior expected length (in cM) of the total genome for which haploid $h_*$ copies from each of $h_1, ..., h_j$, respectively, we calculate the following (let $\Pr_h \equiv \Pr(h_{*(l+1)} \mid X_{l+1} = h_k)$):

$$\begin{aligned} \hat{f}_k \;=\; & \tfrac{1}{\Pr(D)} \textstyle\sum_{l=1}^{L-1} g_l \Big[ \alpha_{kl}\beta_{k(l+1)} \Big[ \exp(-\rho_l/j) + (1.0 - \exp(-\rho_l/j)) f_k \Big] \Pr_h \\ & + (1/2) \Big( \alpha_{kl}\beta_{kl} + \alpha_{k(l+1)}\beta_{k(l+1)} \\ & - 2\alpha_{kl}\beta_{k(l+1)} \Big[ \exp(-\rho_l/j) + (1.0 - \exp(-\rho_l/j)) f_k \Big] \Pr_h \Big) \Big]. \end{aligned} \tag{S30}$$

Unless otherwise noted, in practice we perform initial runs under this protocol where we use 10 steps of E-M to estimate $N_e$ and $\theta$ for each individual in each population. We then average the estimated $N_e$ and $\theta$ across all individuals in the dataset and re-estimate $\hat{f}_k$ for each $k$ using these fixed values.

To make use of our population labels, we then sum these final values of $\hat{f}_k$ across donors $k$ with the same population label. We refer to this vector of $K$ values as the "copying vector" of haploid $h_*$. To get an individual's copying vector, we add the copying vectors of their two haploids. To get a population's copying vector, we average the copying vectors across all individuals within the population. The protocol used to estimate $\hat{f}_k$, e.g. the exact numbers of donors taken from each population, is described in Note S4.1.1.

### A.6 Sampling painted chromosomes

To draw samples of $\vec{X}$ conditional on $(h_*, h_1, ..., h_j, \vec{\rho}, \vec{f}, \theta)$, we perform the following (analogous to (1)), again excluding parameters from conditional probabilities for ease of reading:

1. Sample $X_L$ according to $\Pr(X_L = h_k \mid h_1, ..., h_j, h_*) \propto \alpha_{kL}$.
2. For $l = L-1, ..., 1$, sample $X_l$ according to:

$$\Pr(X_l = h_k \mid h_1, ..., h_j, h_*, X_{l+1}, ..., X_L) \propto \Big[\sum_{i=1}^{j} \alpha_{il}\Big]\Big(1-\exp(-\rho_l)\Big) f_{X_{l+1}} + \alpha_{(X_{l+1})l} \exp(-\rho_l).$$

For each individual in this analysis, we generate 10 such "painting samples" of $\vec{X}$ for each haploid for use in generating coancestry curves (see Note S4.3). Unless otherwise noted, in practice we perform initial runs where we use 10 steps of E-M to estimate $N_e$ and $\theta$ for each recipient individual of a given population label, conditional on the appropriate donor set. We then averaged estimates of $N_e$ and $\theta$ across recipient individuals and fixed these as the values when sampling painted chromosomes.

## References


[1] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure from multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

[2] S.J.E. Baird. Fisher's markers of admixture. *Heredity*, 97:81–83, 2006.

[3] J. E. Pool and R. Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181:711–719, 2009.

[4] P. Moorjani, N. Patterson, J.N. Hirschhorn, A. Keinan, L. Hao, G. Atzmon, E. Burns, H. Ostrer, A.L. Price, and D. Reich. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics*, 7(4):e1001373, 2011.

[5] I. Pugach, R. Matveyev, A. Wollstein, M. Kayser, and M. Stoneking. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol*, 12:R19, 2011.

[6] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotypes data. *Genetics*, 155:945–959, 2000.

[7] D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655–1664, 2009.

[8] D.J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 2012.

[9] Information on materials and methods is available online.

[10] L.L. Cavalli Sforza and Walter Bodmer. *The Genetics of Human Populations*. W.H. Freeman, San Francisco, USA, 1971.

[11] P.R. Loh, M. Lipson, N. Patterson, P. Moorjani, J.K. Pickrell, D. Reich, and B. Berger. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

[12] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:504–509, 2006.

[13] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Gen*, 2(12):e190, 2006.

[14] G.K. Chen, P. Marjoram, and J.D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res*, 19:136–142, 2009.

[15] I. Gronau, M.J. Hubisz, B. Gulko, C.G. Danko, and A. Siepel. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet*, 43:1031–1034, 2011.

[16] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 2011.

[17] R.N. Gutenkunst, R. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10):e1000695, 2009.

[18] A. Keinan, J.C. Mullikin, N. Patterson, and D. Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet*, 39:1251–1255, 2007.

[19] J.N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.*, 128(2):415–423, 2005.

[20] B.M. Henn, L.R. Botigué, S. Gravel, W. Wang, A. Brisbin, J.K. Byrnes, K. Fadhlaoui-Zid, P.A. Zalloua, A. Moreno-Estrada, J. Bertranpetit, C.D. Bustamante, and D. Comas. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet*, 8(1):e1002397, 2012.

[21] G.B.J. Busby, F. Brisighelli, P. Snchez-Diz, E. Ramos-Luis, C. Martinez-Cadenas, M.G. Thomas, D.G. Bradley, L. Gusmo, B. Winney, W. Bodmer, M. Vennemann, V. Coia, F. Scarnicci, S. Tofanelli, G. Vona, R. Ploski, C. Vecchiotti, T. Zemunik, I. Rudan, S. Karachanak, D. Toncheva, P. Anagnostou, G. Ferri, C. Rapone, T. Hervig, T. Moen, J.F. Wilson, and C. Capelli. The peopling of Europe and the cautionary tale of Y chromosome lineage R-M269. *Proceedings of the Royal Society B: Biological Sciences*, 279(1730):884–892, 2012.

[22] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[23] D.M. Behar, B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, E.K. Khusnutdinova, O. Balanovsky, O. Semino, L. Pereira, D. Comas, D. Gurwitz, B. Bonne-Tamir, T. Parfitt, M.F. Hammer, K. Skorecki, and R. Villems. The genome-wide structure of the Jewish people. *Nature*, 466:238–242, 2010.

[24] B.M. Henn, Gignoux C.R., M. Jobin, J.M. Granka, J.M. Macpherson, J.M. Kidd, L. Rodrguez-Botigu, S. Ramachandran, .L. Hon, A. Brisbin, A.A. Lin, P.A. Underhill, D. Comas, K.K. Kidd, P.J. Norman, P. Parham, C.D. Bustamante, J.L. Mountain, and M.W. Feldman. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA*, 108(13):5154–5162, 2011.

[25] B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6):e1000529, 2009.

[26] An online atlas of worldwide human genetic admixture history (available at www.admixturemap.paintmychromosomes.com).

[27] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient Admixture in Human History. *Genetics*, 192(3):1065–1093, 2012.

[28] We convert g inferred generations to the admixture year: 1950 - (g+1) x 28.

[29] J.K. Pickrell and J.K. Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet*, 8:e1002967, 2012.

[30] D. Chandler. *A history of Cambodia*. Westview Press, Boulder, CO, 4th edition, 2007.

[31] C.P. Atwood. *Encyclopedia of Mongolia and the Mongol Empire*. Facts on File, Inc, New York, USA, 2004.

[32] E. E. Bacon. The Inquiry into the History of the Hazara Mongols of Afghanistan. *Southwestern J. Anthropol.*, 7(3):230–247, 1951.

[33] T. Zerjal, Y. Xue, G. Bertorelle, R.S. Wells, W. Bao, S. Zhu, R. Qamar, Q. Ayub, A. Mohyuddin, S. Fu, P. Li, N. Yuldasheva, R. Ruzibakiev, J. Xu, Q. Shu, R. Du, H. Yang, M.E. Hurles, E. Robinson, T. Gerelsaikhan, B. Dashnyam, S.Q. Mehdi, and C. Tyler-Smith. The genetic legacy of the mongols. *Am. J. Hum. Genet.*, 72(3):717–721, 2003.

[34] S. Xu and L. Jin. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *Am. J. Hum. Genet.*, 83(3):322–336, 2008.

[35] M. Gordon. *Slavery in the Arab world*. New Amsterdam Books, New York, USA, 1st edition, 1989.

[36] P. Heather. *Empires and Barbarians: migration, development and the birth of Europe*. Macmillan, London, UK, 2009.

[37] P. Ralph and G. Coop. The geography of recent genetic ancestry across Europe. *PLoS Biology*, 11(5):e1001555, 2013.

[38] C.I. Beckwith. *Empires of the Silk Road: A History of Central Eurasia from the Bronze Age to the Present*. Princeton University Press, Princeton, NJ, 2006.

[39] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2981–2985, 2002.

[40] M Lines. *The Kalasha people of North-western Pakistan*. Emjay Books International, Peshawar, Pakistan, 1999.

[41] R. Chakraborty and K. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Adad. Sci. USA*, 85(23):9119–9123, 1988.

[42] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–33, 2003.

[43] International HapMap Consortium. The International HapMap Project. *Nature*, 426:789–796, 2003.

[44] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–862, 2007.

[45] A. Kong, D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. A high-resolution recombination map of the human genome. *Nat Genet*, 31(3):241–7, 2002.

[46] A.G. Hinch, A. Tandon, N. Patterson, Y. Song, N. Rohland, C.D. Palmer, G.K. Chen, K. Wang, S.G. Buxbaum, E.L. Akylbekova, M.C. Aldrich, C.B. Ambrosone, C. Amos, E.V. Bandera, S.I. Berndt, L. Bernstein, W.J. Blot, C.H. Bock, E. Boerwinkle, Q. Cai, N. Caporaso, G. Casey, L.A. Cupples, S.L. Deming, W.R. Diver, J. Divers, M. Fornage, E.M. Gillanders, J. Glessner, C.C. Harris, J.J. Hu, S.A. Ingles, W. Isaacs, E.M. John, W. H. Kao, B. Keating, R.A. Kittles, L.N. Kolonel, E. Larkin, L. Le Marchand, L.H. McNeill, R.C. Millikan, A. Murphy, S. Musani, C. Neslund-Dudas, S. Nyante, G.J. Papanicolaou, M.F. Press, B.M. Psaty, A.P. Reiner, S.S. Rich, J.L. Rodriguez-Gil, J.I. Rotter, B.A. Rybicki, A.G. Schwartz, L.B. Signorello, M. Spitz, S.S. Strom, M.J. Thun, M.A. Tucker, Z. Wang, J.K. Wiencke, J.S. Witte, M. Wrensch, X. Wu, Y. Yamamura, K.A. Zanetti, W. Zheng, R.G. Ziegler, X. Zhu, S. Redline, J.N. Hirschhorn, B.E. Henderson, H.A. Taylor Jr, A.L. Price, H. Hakonarson, S.J. Chanock, C.A. Haiman, J.G. Wilson, D. Reich, and S.R. Myers. The landscape of recombination in African Americans. *Nature*, 476:170–175, 2011.

[47] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–44, 2006.

[48] A.L. Price, A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519, 2009.

[49] R.R. Hudson, M. Slatkin, and W.P. Maddison. Estimation of levels of gene flow from DNA sequence data. *Genetics*, 132(2):583–589, 1992.

[50] F.M.T.A. Busing, E. Meijer, and R. Van Der Leeden. Delete-m Jackknife for Unequal m. *Statistics and Computing*, 9:3–8, 1999.

[51] G.T. Marth, E. Czabarka, J. Murvai, and S.T. Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166:351–372, 2004.

[52] O. Delaneau, J. Marchini, and J.F. Zagury. A linear complexity phasing method for thousands of genomes. *Nat Methods*, 9(2):179–181, 2012.

[53] O. Delaneau, J.F. Zagury, and J. Marchini. Improved whole chromosome phasing for disease and population genetic studies. *Nat Methods*, 10(1):5–6, 2013.

[54] A. Kong, G. Thorleifsson, D.F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K.T. Kristinsson, S.A. Gudjonsson, M.L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467:1099–1103, 2010.

[55] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*, 81(3):559–575, 2007.

[56] M. Ruhlen. *A Guide to the World's Langauges: Volume I, Classification*. Stanford University Press, Stanford, California, USA, 1991.

[57] S. Xu, W. Huang, J. Qian, and L. Jin. Analysis of Genomic Admixture in Uyghur and Its Implication in Mapping Strategy. *Am. J. Hum. Genet.*, 82(4):883–894, April 2008.

[58] D.W. Anthony. *The Horse The Wheel and Language: How Bronze-Age Riders from the Eurasian Steppes shaped the Modern World*. Princeton University Press, Princeton, US, 2007.

[59] M. Richards, C. Rengo, F. Cruciani, F. Gratrix, J.F. Wilson, R. Scozzari, V. Macaulay, and A. Torroni. Extensive Female-Mediated Gene Flow from Sub-Saharan Africa into Near Eastern Arab Populations. *Am. J. Hum. Genet.*, 72(4):1058–1064, 2003.

[60] N.M. Abu Izzeddin. *The Druzes: a new study of their history, faith and society*. Brill, Leiden, 2nd edition, 1993.

[61] R Elphick, H.B. Gilomee, and J.C. Armstrong. *The Shaping of South African Society, 1652-1840*. Wesleyan University Press, Middletown, Conn. USA, 1st edition, 1989.

[62] J. Diamond and P. Bellwood. Farmers and their languages: The first expansions. *Science*, 300(5619):597–603, 2003.

[63] J.K. Pickrell, N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, T. Guldemann, B. Kure, S.W. Mpoloka, H. Nakagawa, C. Naumann, M. Lipson, P. Loh, J. Lachance, J. Mountain, C.D. Bustamante, B. Berger, S.A. Tishkoff, B.M. Henn, M. Stoneking, D. Reich, and B. Pakendorf. The genetic prehistory of southern Africa. *Nat Commun*, 3(1143), 2012.

[64] P. Mitchell. *The archaeology of Southern Africa*. Cambridge University Press, Cambridge, UK, 2002.

[65] P. Mitchell. Genetics and southern African prehistory: An archaeological view. *JAS*, 88:73–92, 2010.

[66] S Wolpert. *A new history of India*. Oxford University Press, New York, 2000.

[67] J.S. Olson. *An Ethnohistorical Dictionary of China*. Greenwood Publishing Group, Westport, CT, USA, 1998.

[68] L Fengxiang and L Whaley. The world loanword database. http://wold.livingsources.org/language/20.

[69] S.A. Tishkoff, F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J.B. Hirbo, A.A. Awomoyi, J.M. Bodo, O. Doumbo, M. Ibrahim, A.T. Juma, M.J. Kotze, G. Lema, J.H. Moore, H. Mortensen, T.B. Nyambo, S.A. Omar, K. Powell, G.S. Pretorius, M.W. Smith, M.A. Thera, C. Wambebe, J.L. Weber, and S.M. Williams. The Genetic Structure and History of Africans and African Americans. *Science*, 324(5930):1035–1044, 2009.

[70] C. Ehret and M. Posnansky. The first spread of food production to Africa. In *The Archaeological and Liguistic Reconstruction of African History*. Berkeley: University of California Press, California, USA, 1982.

[71] M Maceachern. Two thousand years of West African History. In A. B. Stahl, editor, *African Archaeology*. Blackwell, Oxford, UK, 2005.

[72] V. Lieberman. *Strange Parallels: South-East Asia in a Global Context: 800-1830*, volume 1: Integration with the mainland. Cambridge University Press, Cambridge, UK, 2003.

[73] R. Chaix, C. Cao, and P. Donnelly. Is Mate Choice in Humans MHC-Dependent? *PLoS Genet*, 4(9):e1000184, 2008.

[74] J.K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001.

[75] A.J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29(2):217–22, 2001.

[76] D.C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M.J. Rieder, D.A. Nickerson, and M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet*, 36(7):700–6, 2004.

[77] G.A. McVean, S.R. Myers, S. Hunt, P. Deloukas, D.R. Bentley, and P. Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(5670):581–4, 2004.

[78] P. Fearnhead and N.G. Smith. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am J Hum Genet*, 77(5):781–94, 2005.

[79] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–4, 2005.

[80] J.L. Newman. *The peopling of Africa*. Yale University Press, USA, 1995.

[81] L. Pagani, T. Kivisild, A. Tarekegn, R. Ekong, C. Plaster, I. Gallego-Romero, Q. Ayub, S.Q. Mehdi, M.G. Thomas, D. Luiselli, E. Bekele, N. Bradman, D.J. Balding, and C. Tyler-Smith. Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool. *The American Journal of Human Genetics*, 91(1):83–96, 2012.

[82] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*, 77:257–286, 1989.